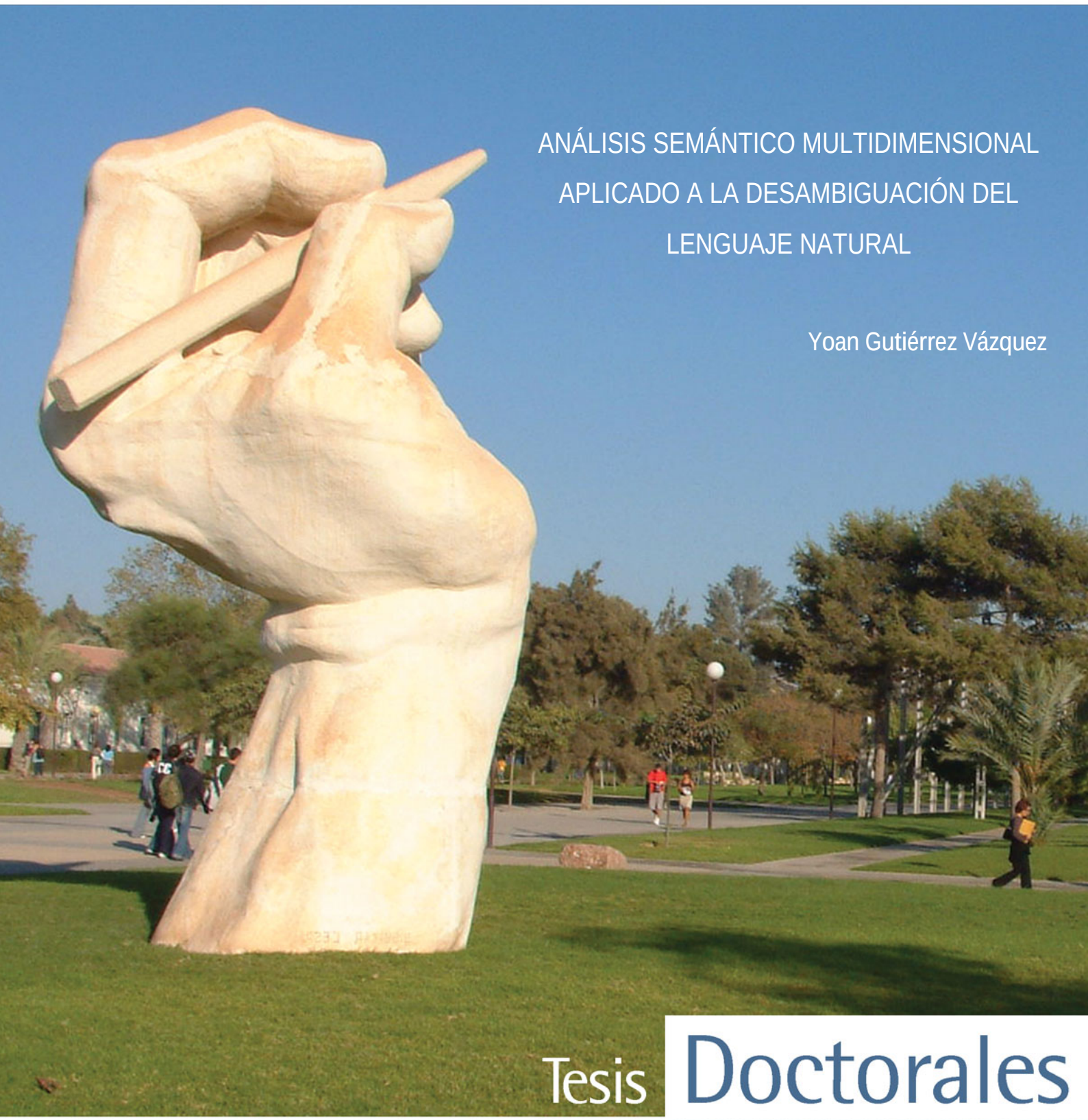
Universitat d'Alacant
Universidad de Alicante

ANÁLISIS SEMÁNTICO MULTIDIMENSIONAL APLICADO A LA DESAMBIGUACIÓN DEL LENGUAJE NATURAL

Yoan Gutiérrez Vázquez



Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

ANÁLISIS SEMÁNTICO MULTIDIMENSIONAL APLICADO A LA DESAMBIGUACIÓN DEL LENGUAJE NATURAL

Tesis Doctoral

Autor: Yoan Gutiérrez Vázquez

Directores: Dr. Andrés Montoyo Guijarro

Dra. Sonia Vázquez Pérez

Depto. de Lenguajes y Sistemas Informáticos

Universidad de Alicante

Alicante, Enero del 2012

A mis princesas...

AGRADECIMIENTOS

En primer lugar le quiero agradecer a mi tutor Andrés Montoyo, por haber confiado en mí desde el primer momento, por su apoyo incondicional tanto intelectual como fraternal. Él sin duda alguna, me ha mostrado una visión de la ciencia muy diferente a como yo la había percibido, sin su ayuda ninguno de los logros que hemos alcanzado en esta investigación serían posibles. Por todo lo que ha representado para mí, le estaré eternamente agradecido.

Qué puedo decir de mi tutora Sonia Vázquez, si ella no tuvo reparo alguno en tomar la decisión de colaborar juntos, sin tan siquiera conocerme. Ella que ha asumido mi ritmo de trabajo redactando juntos tantos artículos científicos, con grandes presiones de tiempo. Que nunca ha evadido ninguna de las ayudas que le he pedido. Ella que como digo yo, siempre está al pie del cañón para responder en tiempo. A ella, le agradezco inmensamente su gran ayuda.

No pueden faltar los reconocimientos familiares y fraternales. Quizás deba parte de esta Tesis a mi esposa Lianne, que ha sabido comprender la importancia de tanto sacrificio. Que ha asumido heroicamente su rol como esposa, compañera y mejor madre, para de esta forma darme fuerzas y seguir adelante con mis sueños. Ella, que ha vivido cada uno de mis momentos de agotamiento y siempre ha estado para animarme. A ella, que es una de mis princesas, le estoy inmensamente agradecido.

Aunque parezca extraño, a mi princesita Melissa también le agradezco. Porque sin temor a equivocarme, es lo mejor que me ha pasado en la vida. Si en la elaboración de toda esta jornada científica se necesitaba energías, esas me las ha dado mi princesita. Con solamente uno de sus gestos obtengo el estímulo necesario para seguir adelante.

A mi gran madre, mi reina, que vive y sufre todos mis pesares. Que se siente muy orgullosa de mí al igual que mi padre. Ellos que, han sabido formar a sus hijos del mejor modo posible y que en momentos como estos reciben el mejor de los premios. A ellos les debo todo lo que soy y lo que seré.

No puedo dejar de dar las gracias a Tony, que siempre responde al llamado de un amigo, gracias por tus consejos que siempre he considerado de gran importancia, gracias por todo.

Son muchas las personas a las que debo agradecer, pues considero importante desde el que pregunta "cómo va la Tesis", hasta el que pasa un momento de distracción conmigo. Todas esas pequeñas cosas, para mí son muy grandes.

Le agradezco a Dios, por darme fuerzas……

ÍNDICE GENERAL

1. Introducción ..... - 1 -

1.1. Arquitectura de los Sistemas de PLN ..... - 2 -

1.2. Problemática del PLN ..... - 3 -

1.2.1. Ambigüedad del Lenguaje ..... - 4 -

1.3. Desambiguación del Sentido de las Palabras ..... - 4 -

1.4. Aplicaciones ..... - 6 -

2. Estado del Arte ..... - 11 -

2.1. Introducción ..... - 11 -

2.2. Proceso de selección de sentidos correctos ..... - 12 -

2.3. Contexto de la palabra ..... - 14 -

2.4. Fuentes de conocimiento externas ..... - 16 -

2.4.1. Lenguajes de representación del conocimiento ..... - 20 -

2.4.2. WordNet ..... - 21 -

2.4.3. Recursos de integración semántica ..... - 30 -

2.5. Métodos de clasificación ..... - 31 -

2.5.1. WSD con supervisión ..... - 31 -

2.5.2. WSD débilmente supervisado ..... - 32 -

2.5.3. WSD sin supervisión ..... - 32 -

2.5.4. Métodos basados en conocimiento ..... - 34 -

2.6. Evaluación de la desambiguación ..... - 46 -

2.6.1. SENSEVAL. *Evaluation Exercises for the Semantic Analysis of Text* ..... - 46 -

2.7. Conclusiones ..... - 59 -

3. Integración de Recursos Semánticos ..... - 63 -

3.1. Introducción ..... - 63 -

3.2. Primera fase de integración ..... - 63 -

3.2.1. Propuesta de integración ..... - 65 -

3.2.2. *Software* y librerías de clases ..... - 65 -

3.3. Segunda fase de integración ..... - 67 -

3.3.1. Propuesta de integración ..... - 67 -

3.3.2. Librerías de clases ..... - 69 -

3.4. Análisis de resultados ..... - 70 -

3.4.1. Resultados de la primera fase de integración ..... - 70 -

3.4.2. Resultados de la segunda fase de integración ..... - 72 -

3.5. Investigaciones que han utilizado ISR-WN ..... - 73 -

3.6. Conclusiones ..... - 74 -

4. Aproximaciones de Resolución de Ambigüedad Semántica basadas en el Análisis Semántico Multidimensional ..... - 77 -

4.1. Introducción ..... - 77 -

4.2. Aproximaciones Basadas en Árboles Semánticos ..... - 78 -

4.2.1. Árboles Semánticos Relevantes ..... - 78 -

4.2.2. Árboles Semánticos Relevantes combinados con Frecuencias de Sentidos ..... - 92 -

4.3. Aproximaciones Basadas en Grafos ..... - 95 -

4.3.1. Aplicación de técnicas de *N-Cliques* ..... - 95 -

4.3.2. *PageRank* combinado con Frecuencias de Sentidos ..... - 102 -

4.4. Evaluaciones y resultados ..... - 107 -

4.4.1. Árboles Semánticos Relevantes ..... - 107 -

4.4.2. Árboles Semánticos Relevantes combinados con Frecuencias de Sentidos ..... - 110 -

4.4.3. *N-Cliques* combinado con *Reuters Vector* ..... - 112 -

4.4.4. *N-Cliques* combinado con Árboles Semánticos Relevantes ..... - 115 -

4.4.5. *Personalizing PageRank* combinado con Frecuencias de Sentidos basado Múltiples Dimensiones ..... - 117 -

4.4.6. Comparaciones con propuestas novedosas ..... - 122 -

4.4.7. Evaluación general ..... - 124 -

4.5. Conclusiones ..... - 127 -

5. Aplicaciones del Análisis Semántico Multidimensional en otras tareas del Procesamiento del Lenguaje Natural ..... - 129 -

5.1. Introducción ..... - 129 -

5.2. Procesamiento de opiniones basado en el análisis de múltiples dimensiones semánticas ..... - 131 -

5.2.1. Obtención de RST's de conceptos ..... - 132 -

5.2.2. Obtención de Árboles Semánticos de Polaridad Positiva ..... - 132 -

5.2.3. Obtención de Árboles Semánticos de Polaridad Negativa ..... - 133 -

5.2.4. Obtención de polaridades de las frases ..... - 134 -

5.2.5. Detección de frases que contienen opiniones ..... - 135 -

5.2.6. Determinación de frases relevantes a preguntas ..... - 135 -

5.2.7. Clasificación de la polaridad de las frases ..... - 136 -

5.3. Clasificación de textos basado en el análisis de múltiples dimensiones semánticas ..... - 136 -

5.3.1. Obtención de RST's de textos y categorías ..... - 137 -

5.3.2. Normalización de Vectores ..... - 138 -

5.3.3. Aplicación de técnicas de clasificación ..... - 138 -

5.4. Evaluaciones y resultados ..... - 147 -

5.4.1. Evaluación del análisis de opiniones con respecto a la detección de opiniones, relevancia y polaridad ..... - 147 -

5.4.2. Evaluación de la clasificación de textos aplicado a opiniones ..... - 150 -

5.5. Conclusiones ..... - 155 -

6. Conclusiones y trabajos futuros ..... - 157 -

6.1. Conclusiones generales ..... - 157 -

6.2. Trabajos futuros ..... - 158 -

6.3. Producción científica ..... - 159 -

Acrónimos ..... - 161 -
Referencias Bibliográficas ..... - 163 -

Según vamos adquiriendo conocimiento, las cosas no se hacen más comprensibles, sino más misteriosas.

Albert Schweitzer, 1875-1965. Teólogo, filósofo, médico, escritor, músico y misionero alemán.
Premio Nobel de la Paz en 1952.

1. INTRODUCCIÓN

En la actualidad, la gran explosión de tecnologías ha motivado el desarrollo parejo de diferentes técnicas estrechamente vinculadas a mejorar la comunicación hombre-máquina. La aparición de Internet unida a las nuevas tendencias de comunicación vía mensajes cortos, participaciones en foros, redes sociales, etc., ha revolucionado la forma en que las personas se comunican, trabajan e incluso gestionan su tiempo libre. Como consecuencia de esta revolución tecnológica se generan grandes cantidades de información sobre diferentes formatos, temáticas y ámbitos sociales. Estos grandes volúmenes de información presentados hoy en día en su mayoría en Internet mediante archivos documentales, *fórums*, *blogs*, *microblogs* y redes sociales de comunicación han generado grandes expectativas en la forma en que las personas se comunican, y comparten conocimiento y emociones; además de influir en el comportamiento social, político y económico mundial.

A partir de estos hechos, se hace evidente la necesidad de gestionar de algún modo esta información. Por ejemplo, las búsquedas en Internet proporcionan infinidad de resultados que han de ordenarse según diferentes criterios, se debe determinar el idioma de las búsquedas, etc. También las traducciones de textos se hacen de forma automática con un software especialmente diseñado para ello y otras tareas que el hombre ha tomado como necesidades para la vida diaria. Como también el uso de herramientas para el análisis de noticias y opiniones, que orientan el curso en que los internautas enfocan y el perciben el pasado, presente y el futuro.

Se refiere entonces, a la necesidad de procesar el lenguaje humano no únicamente para los ejemplos antes mencionados, sino para otras muchas tareas como: la corrección de documentos, traducción automática, elaboración de resúmenes, extracción y valoración de opiniones, etc. Todas estas tareas requieren de un profundo conocimiento lingüístico y a la vez, en muchos casos, de un elevado coste computacional. Las mencionadas necesidades derivadas de la aparición de nuevas tecnologías de la información han generado una disciplina denominada 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. El PLN se ha convertido en una rama informática dentro del campo de la Inteligencia Artificial (en inglés *Artificial Intelligence* (AI)), la cual se puede definir como "la ciencia y la ingeniería que hace las máquinas inteligentes" (McCarthy, 1959). Logrando entonces con este subconjunto de la AI que se estudie y analice el modo de procesar lenguaje natural, tal y como el ser humano lo genera, mediante mecanismos y recursos informáticos.

Entre los mayores impulsores que ha propiciado el desarrollo del PLN se encuentra Internet, donde la información aparece en su mayoría descentralizada y desestructurada. Estas problemáticas han generado la necesidad de poseer herramientas útiles para la gestión de toda la información de modo rápido y eficaz. Lo que propicia la creación de dos líneas de investigación bien diferentes y definidas: Tecnologías de Procesamiento de Datos (TPD) y Tecnologías de Procesamiento del Lenguaje Natural (TPLN); donde cada una procesa de forma diferente la información. A diferencia de las TPD que se enfoca en reducir el espacio ocupado, almacenar de forma óptima los datos, disminuir tiempos de respuesta en la búsqueda, etc., las TPLN requieren un conocimiento profundo del lenguaje para poder procesar la información textual no estructurada y emitir resultados fiables.

A diferencia de las tecnologías de procesamiento de datos los sistemas de PLN deben profundizar todavía más en el conocimiento lingüístico para su correcto funcionamiento. Por ejemplo, un sistema de traducción automática debería ser capaz de traducir correctamente al idioma inglés la oración: Juan ha muerto y entonces él irá al cielo. El resultado con relación a la palabra cielo podría ser entre otros los que siguen a continuación.

- *Juan is dead and he will go to **sky**.*
- *Juan is dead and he will go to **heaven**.*

Como se puede observar el uso de la palabra ***sky*** (el espacio en el que se mueven los astros y por efecto visual parece rodear la Tierra) y ***heaven*** (lugar de morada de los dioses, ángeles y almas humanas) indican notables desviaciones del significado de la palabra en traducción. Como consecuencia, cuando se necesita un conocimiento más preciso del lenguaje, de las relaciones entre palabras o de las expresiones en diferentes contextos, se requiere de PLN.

Todo sistema de Procesamiento del Lenguaje Natural requiere amplio conocimiento sobre las estructuras del lenguaje. Para afrontar su análisis se enfocan diferentes niveles de comprensión, como pueden ser el fonológico léxico[1], morfológico[2], sintáctico[3], semántico[4], pragmático[5]. En cada nivel se procesa el texto de forma distinta, pretendiendo extraer un tipo de información determinada, desarrollando y requiriendo distintos recursos y técnicas. A continuación se describe la arquitectura de un sistema de PLN.

1.1. ARQUITECTURA DE LOS SISTEMAS DE PLN

Los sistemas de PLN, de modo general, presentan tres módulos principales, de análisis léxico, sintáctico y semántico (Vázquez, 2009, Montoyo, 2002).

- **Léxico:** Su objetivo es detectar la menor unidad de caracteres posibles que denoten significado (palabras). Estas palabras pueden ser simples, compuestas, siglas, etc. Es necesario diferenciar entre la forma (tal y como se encuentra la palabra en el texto) y el lema (la forma canónica de la palabra). Como resultado se debe asociar cada palabra con su lema, detectar sus posibles categoría gramaticales, género, número, etc.
- **Sintáctico:** A través del análisis sintáctico se estructura la frase (sujeto, predicado (sustantivo, adjetivo, verbo, adverbio)) y se selecciona la etiqueta gramatical más apropiada para cada palabra.
- **Semántico:** En este caso se ocupa de asignar el sentido correspondiente a cada palabra. Existen diferentes técnicas aplicables para resolver la ambigüedad: lógica de predicados, redes semánticas, grafos de dependencias conceptuales, etc.

Como se observa en la Figura 1 de modo general, el proceso se inicia tras introducir el texto, después se aplica el análisis léxico obteniendo las palabras, utilizando diccionarios u otras fuentes de información. En el siguiente paso interviene el módulo de análisis sintáctico y en caso de necesitar más información se procede al semántico. Los subprocesos implicados tras el análisis léxico suelen ayudarse de varias fuentes de información, como son: los diccionarios, lexicones, tesauros, bases de datos con información sobre la gramática del lenguaje, ontologías y otras.

[1] La fonología léxica da cuenta de las interacciones de la morfología y la fonología en el proceso de construcción de palabras.

[2] La morfología es la identificación, análisis y descripción de la estructura de los morfemas y otras unidades de significado en un lenguaje como palabras, afijos, las partes del discurso, etc.

[3] La sintaxis es el estudio de los principios y normas para la construcción de frases y oraciones en las lenguas naturales.

[4] La lingüística semántica se centra en la identificación del significado que denotan las palabras en el contexto de la expresión humana a través del lenguaje.

[5] La pragmática estudia las formas en que el contexto contribuye a su significado. El conocimiento pragmático ayuda a interpretar la oración dentro de su contexto.

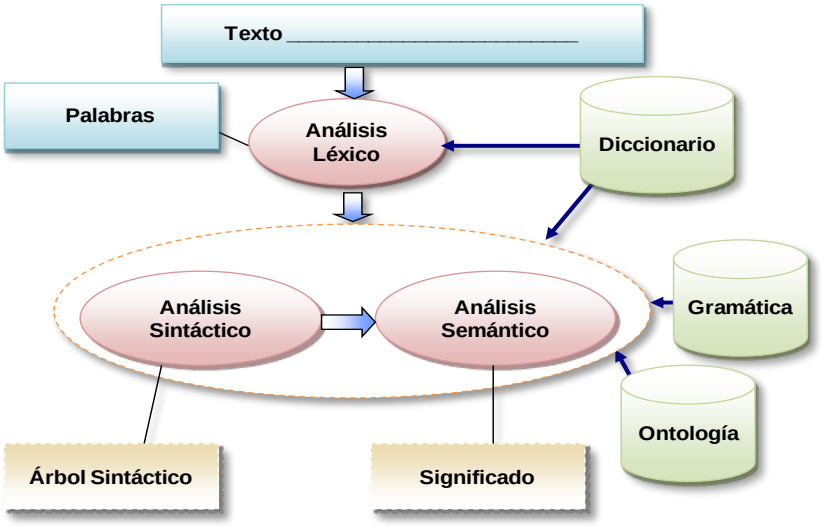


Figura 1. Estructura general de un sistema de PLN.

1.2. PROBLEMÁTICA DEL PLN

La puesta en práctica de todas las formas de conocimiento lingüístico que intervienen en la ejecución de los sistemas de PLN evidencia la aparición de un problema común que resolver, las ambigüedades del lenguaje. Para ser capaces de crear sistemas eficientes de Procesamiento del Lenguaje Natural se hace necesario convertir la información que ofrecen las palabras del texto plano, en palabras manipulables a nivel de Conceptos. Por ejemplo para la palabra *plant* del idioma inglés:

- *I have a **plant** sown at the garden.*
- *Brad is going to **plant** an apple tree at the park.*
- *All people on my family are working in this industrial **plant**.*

La palabra *plant* tiene diferentes significados en las tres frases de ejemplo. Entre la primera y la segunda frase la categoría gramatical es la responsable de distinguir su significado. Y entre las dos primeras y la tercera, el contexto asociado es el encargado de proporcionar su significado. Por ejemplo, el dominio *Botanic* y el dominio *Industry* respectivamente. Por consiguiente, se puede afirmar que detectar el significado correcto de las palabras no representa una tarea fácil (Gutiérrez *et al.*, 2011b). La identificación del sentido correcto de las palabras según el contexto donde estas se utilicen es llamada comúnmente *Word Sense Disambiguation* (WSD).

Por este motivo, al diseñar un sistema de PLN uno de los objetivos fundamentales es resolver las múltiples ambigüedades que puede presentar el lenguaje o lo que es igual, realizar un **análisis semántico** del texto. Las ambigüedades del lenguaje pueden ser de diferentes tipos: estructural, léxica, de ámbito de cuantificación, de función contextual y referencial (Montoyo Guijarro, 2002, Mezquita, 2008). Se pueden encontrar expresiones y palabras que consiguen tener significados distintos dependiendo en qué circunstancias se usen. 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) y ambigüedad (pudiendo darse diversos significados de una palabra o una frase según el contexto).

El conocimiento del lenguaje natural viene asociado a la difícil tarea de resolver sus diferentes ambigüedades. Al diseñar un sistema de PLN se deben encontrar las posibles soluciones, y seleccionar aquella que resuelva la ambigüedad.

1.2.1. AMBIGÜEDAD DEL LENGUAJE

Las ambigüedades del lenguaje se pueden apreciar desde diferentes perspectivas. Existen ambigüedades debidas a palabras polisémicas o ambigüedades producidas por las distintas interpretaciones que pueda tener una oración. Los diferentes tipos de ambigüedad se pueden clasificar de la siguiente manera:

- **Ambigüedad léxica:** Tal y como (Vázquez, 2009) muestra, 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 componentes de una oración, se puede interpretar de varias formas. Siendo a veces casi imposible de solucionar.
  Por ejemplo: Pedro vio a su abuela sin usar gafas (¿Pedro no usó sus gafas para ver a su abuela o Pedro vio que su abuela no tenía las gafas colocadas?)
- **Ambigüedad semántica:** Dentro de este tipo de ambigüedad se pueden diferenciar tres clases:
  - Ambigüedad debida a las **palabras polisémicas**. En este caso, una misma palabra puede tener distintos significados dependiendo de su uso.
    - La **ejecución** del recluso ya está programada. (Ejecución como eliminación física de una persona condenada a muerte)
    - El software está en **ejecución**. (Poner a funcionar un programa de computadora)
  - Ambigüedad debida a encontrar **una misma estructura sintáctica con diferentes significados**.
    - Todos los estudiantes de secundaria hablan dos lenguas. (¿Cada estudiante habla dos lenguas o exclusivamente se hablan dos lenguas determinadas?)
  - **Ambigüedad referencial**. Para este caso el análisis debe exceder los límites de la frase. Para poder conocer en textos previos o futuros los antecedentes referenciales de los pronombres.
    - ¡El pollo está en el horno, ábrelo ya! (¿Hacen referencia al pollo o al horno?).

Tras haber observado estos ejemplos, es evidente el por qué la problemática de la resolución de ambigüedad ha sido por muchos años una vertiente a estudiar, por tal motivo la resolución de distintos tipos de ambigüedades deriva la temática denominada Desambiguación del Sentido de las Palabras.

1.3. DESAMBIGUACIÓN DEL SENTIDO DE LAS PALABRAS

WSD es considerada una **tarea intermedia** como lo es también la **tokenización**[6], el **análisis morfológico**[7], el **análisis sintáctico**[8] y el **reconocimiento de entidades**[9]. WSD consiste en establecer de las definiciones de una palabra, cuál es la más adecuada según el contexto. Este tipo de técnicas se basan en identificar una serie de características comunes entre el contexto y

[6] Divide el texto en componentes simples como palabras, números, signos de puntuación.
[7] Extrae la categoría gramatical, rasgos sobre género, número y persona de las palabras de un texto.
[8] Obtiene sintagmas y relaciones estructurales (sintácticas) entre ellos.
[9] Reconoce conjuntos de palabras que hacen referencia a una entidad como por ejemplo: el nombre de una persona, empresa o localidad, una fecha, una cantidad.

las definiciones de la palabra, para luego asignar la definición más probable como sentido correcto (Izquierdo, 2010).

Estas tareas intermedias generan información directamente útil para el usuario y se denominan **aplicaciones finales**. Como son, la **recuperación de información** (permite obtener un subconjunto de documentos relevantes para el usuario a partir de un conjunto mucho mayor), la **extracción de información** (extrae información relevante de un documento que contiene texto no estructurado (lenguaje natural)), la **búsqueda de respuestas** (devuelve respuestas a preguntas realizadas por el usuario en lenguaje natural), la **traducción automática** (traduce un texto en lenguaje natural de un idioma a otro), la **minería de opiniones** (extrae información relevante de documentos asociado a detección de opiniones, detección de polaridades de estas, detección de emociones, etc. **)** entre otras.

Por lo general las tareas intermedias y finales funcionan en cadena (la salida de una de ellas es la entrada para otra), es decir, hay tareas que necesitan la información que ofrecen otras para realizar su labor. La utilidad real de las tareas intermedias está en su uso dentro de una aplicación final. Coincidiendo entonces con (Ide and Véronis, 1998) al considerar que WSD es una tarea esencial para aplicaciones de PLN. Dado la importancia que esta representa, se deben considerar algunos de los factores que provocan el incremento de la complejidad y en consecuencia aumento de tiempo de ejecución y variación en la efectividad. Los factores principales según afirma (Navigli, 2009) son:

- La tarea puede **ejecutarse** sobre **textos orientados a dominios** en particular (indican que el contenido del texto se encuentra relacionado en un área en particular) **o abiertamente** implicando informaciones sumamente variables.
- La **granularidad de los sentidos** (etiquetas o conceptos) que serán asignados a las palabras amplía o restringe el universo de posibilidades a manipular.
- Los **sentidos** pueden estar **previamente representados** como base de conocimiento o pueden generarse dinámicamente.
- La tarea puede **implicar** a unas **pocas palabras o a todas** las que formen parte de la frase.

Por otra parte se pueden encontrar otros factores asociados al conocimiento del que ayudan los sistemas de WSD. Por ejemplo:

- Para un conjunto de palabras que denotan una frase, el sistema de WSD tiene que ser capaz de obtener de las **fuentes de conocimiento,** las herramientas necesarias para determinar lograr obtener el sentido correcto para cada palabra objetivo. Según técnica a utilizar estas fuentes de están sujetas a variar. Las que pueden ser ejemplos de corpus anotados a nivel de sentido o no y también recursos estructurados como pueden ser diccionarios, tesauros, redes semánticas y otros.

Para llevar a cabo la tarea de WSD se necesita al menos una fuente de conocimiento, lo que significa que existe una necesidad de construir dichas fuentes de forma que sean útiles para sistemas de WSD. La creación de nuevas fuentes de conocimiento requiere del esfuerzo de especialistas del lenguaje y el consumo de mucho tiempo y personal especializado. Además, cada idioma demanda un tratamiento personalizado y fuentes de conocimiento propias. En el análisis de WSD se tienen en cuenta recursos como:

- El contexto de la palabra
- Recursos de conocimiento (recursos léxicos, redes semántica, ontologías, etc.)

A continuación se describen diversas aplicaciones de PLN donde es necesario el uso de sistemas de WSD para mejorar los resultados obtenidos.

1.4. APLICACIONES

Son muchas las tareas que necesitan del análisis semántico para lograr su buen funcionamiento. Aunque la mayoría de ellas no la usan debido a los bajos resultados logrados hoy en día en la tarea de WSD y por consiguiente no desean introducir ruidos en sus sistemas. En esta sección se explican las principales razones por las que en esta Tesis se ha decidido investigar en la generación y uso de la información semántica. Las investigaciones que se mencionan a continuación evidencian el uso que por parte de diferentes tareas de PLN hacen uso del análisis semántico.

El proceso WSD en muchas ocasiones se presenta unido a otros que forman parte de la disciplina de PLN, según se plantea en (Wilks and Stevenson, 1996, Wilks and Stevenson, 1998) usualmente es considerado un proceso intermedio necesario en otras tareas llamadas finales, como son la traducción automática, recuperación de información, entre otros. Brevemente se reflejará cuán importante es este proceso intermedio en alguna de las temáticas nombradas, como por ejemplo:

- **Recuperación de Información:** El proceso de desambiguación ayuda a determinar qué documentos contienen las palabras de entrada que coinciden en sentido con las recuperadas. Incluso es importante discernir al menos que la información buscada corresponde contextualmente con la recuperada. Varios autores (Andrews and Rajman, August 2004, Vilares, 2005, Padilla, 2002b, Santiago, 2004, Padilla, 2002a) plantean que la discriminación otorgada mediante la desambiguación, es vital para la fiabilidad de las respuestas de un sistema de recuperación de información.
- **Clasificación de Documentos:** Para determinar la temática de un texto o fragmento de texto es imprescindible conocer exactamente qué quiere decir cada una de las palabras contenidas en el texto, lo que hace necesario conocer su significado. Para ello (Schneider, 2005, HinrichSchütze, 1997) consideran el uso de la clasificación con la visión de sostener que una palabra se define por su contexto y un contexto individual es a su vez se define por sus propios contextos. Entonces los similares contextos individuales pertenecen a la misma palabra, conceptualizando de esta manera a las palabras o grupos de palabras. Esta visión semántica más bien identifica documentos y palabras fuertemente relacionadas revelando internamente su semántica.
- **Búsquedas de Respuestas:** obviamente comprender el significado correcto de una pregunta es crucial para poder obtener una respuesta correcta. Conocer el sentido de una palabra permite obtener sinónimos de ella, que pueden ser cruciales para adquirir la respuesta. Por ejemplo, ante la pregunta:
    - ¿Quién creó el teléfono?, la respuesta solamente se puede conseguir a partir del fragmento de texto "Alexander Graham Bell. Este construyó el primer teléfono en 1876, en el estado de *Massachussets*, en los Estados Unidos."[10], si fuéramos capaces de determinar que en este caso crear e inventar son sinónimos.
- **Traducción Automática:** WSD se cataloga sumamente importante en los procesos de traducción automática (Hedlund et al., 2001, Kwok, 1999, Boguslavsky et al., 2005, Clough, 2005). Son múltiples las traducciones que pueden tener las palabras, por lo que el contexto en que se encuentra es vital en la exactitud de las respuestas, por ejemplo la palabra en inglés *disc* tiene varias traducciones al español.

[10] http://www.misrespuestas.com/quien-invento-el-telefono.html

- disco de lanzamiento
- disco de grabación de vinilo
- dispositivo de memoria o disco magnético
- plato circular aplastado

Según (Warren, 1955) si se analiza cada palabra individualmente para obtener su significado, sería imposible, pero si el análisis es en conjunto con las otras se podría decidir.

- **Análisis de Opiniones:** Según afirma (Martín *et al.*, 2010) sus estudios sobre la influencia de la determinación semántica de las opiniones proyecta resultados positivos y negativos. Pues cuando se tiene controlada la tarea de WSD los resultados son realmente alentadores, así ha sido demostrado en dicha investigación. Sin embargo, si WSD constituye una tarea intermedia en la clasificación de la polaridad, los errores de desambiguación podría afectar a la calidad de clasificación. Esto proporciona una motivación adicional para estudiar en profundidad este problema, debido a la falta de corpus anotados manualmente con todos los sentidos y la polaridad. Además, el algoritmo de **desambiguación depende de los recursos utilizados y el conocimiento**.

Como línea general del estudio de emociones y opiniones se encuentra la **Computación Afectiva** (conocida en inglés como *Affective Computing* (AC)). Esta ha estado condicionada por la influencia creciente y decisora de las Web Sociales (web de interacción social y comunicación). La AC se soporta sobre la base del análisis de texto generado sobre la web (a través de *blogs*, *forums*, *wikis*, sitios de revisión o *microblogs*). Estos textos expresados sobre la web generan opiniones, comentarios y noticias que tienen lugar en la sociedad. Los grandes volúmenes de información subjetiva presentes en internet han producido una peculiar manera en la que las personas se comunican, comparten información y emociones. Transformándose esta nueva realidad, donde las noticias y sus opiniones al respecto circulan, dando lugar a fenómenos sociales, económicos y psicológicos nuevos y desafiantes. Acerca de ese fenómeno y orientado al estudio de la extracción de cruciales conocimientos que diariamente están contenidos en las fuentes de opiniones, nuevas áreas de investigación se han creado. Como son la **detección de subjetividad**, la **extracción** y **clasificación de opiniones** (positiva, negativa y neutral); todas estas pertenecientes a la AC (Picard, 1995). Principalmente las AC abordan el análisis de la subjetividad (que trata de "estados privado" (Banfield, 1982), un término que encierra los sentimientos, opiniones, emociones, valoraciones, creencias y especulaciones), el análisis de los sentimientos y la minería opinión (Pang and Lee, 2003). En el análisis de subjetividad se incluye la clasificación de los textos de acuerdo con la emoción expresada.

Tradicionalmente no se ha obtenido grandes mejoras en aplicaciones finales del PLN, con el uso de información semántica (Vázquez, 2009). Se considera que los principales problemas radican en que no se conoce la forma óptima de integrar dicha información y por otro lado, los mejores sistemas actuales de WSD no superan el 69% de precisión y cobertura para la tarea de todas las palabras (*All Words*[11]). Esto indica, que no se dispone de garantías para el uso confiable de sistemas intermedios del PLN. A pesar que los resultados alcanzados son bajos, los mejores se han obtenido por sistemas que requieren de entrenamiento. Para estos sistemas es necesario tener de antemano grandes cantidades de información anotada y sus respuestas dependen de los dominios sobre del aprendizaje. Sin ligaduras de ningún tipo, es posible encontrar los sistemas sin supervisión, pero hasta la actualidad sus aciertos no sobrepasan el 59% de precisión y cobertura para la tarea de todas las palabras. Quizás, estos resultados estén ligados a que normalmente los sistemas de WSD utilizan solamente una base de conocimiento,

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

pudiendo servirse de varias. Por esta razón, desarrollar métodos informáticos sin supervisión que sean más confiables resultaría de mucha ayuda para el desarrollo de la temática. Después de haber hecho estos análisis surge el siguiente **problema**. ¿Es posible con la aplicación no supervisada y basada en conocimiento del Análisis Semántico Multidimensional en la Resolución de Ambigüedad Semántica de las Palabras, superar los resultados de sistemas actuales de similar condición y además poder aplicar este análisis en el área de la Minería de Opiniones?

Se propone la siguiente **hipótesis:** Con la aplicación no supervisada y basada en conocimiento del Análisis Semántico Multidimensional en la Resolución de Ambigüedad Semántica de las Palabras, es posible superar los resultados de sistemas actuales de similar condición y además se podrá aplicar este análisis en el área de la Minería de Opiniones.

El **objetivo principal** de esta Tesis es el desarrollo de recursos y métodos informáticos no supervisados y basados en conocimiento, soportados sobre la base de la Semántica Multidimensional, capaces en conjunto, de superar los resultados alcanzados por sistemas sin supervisión de Resolución de la Ambigüedad Semántica de las Palabras y además aplicarlos en el área de la Minería de Opiniones. Con ese propósito, se propone **el estudio** de la Semántica Multidimensional, para **aplicarla en** la Resolución de la Ambigüedad Semántica de las Palabras y la Minería de Opiniones, en concreto para el idioma inglés.

El trabajo se centra en los siguientes objetivos específicos:

- Estudiar la temática asociada a la desambiguación del sentido de las palabras y los trabajos precedentes.
- Desarrollar una herramienta capaz de integrar recursos semánticos para poder aplicar Análisis Semántico Multidimensional en tareas del Procesamiento del Lenguaje Natural.
- Desarrollar métodos informáticos no supervisados que apliquen un Análisis Semántico Multidimensional en la tarea de Resolución de la Ambigüedad Semántica de las Palabras.
- Analizar la influencia del Análisis Semántico Multidimensional en el área de la Minería de Opiniones.
- Evaluar los recursos y los métodos.

La Tesis de divide en los siguientes capítulos:

Seguido de la introducción se presenta el Capítulo 2, donde se abordan conceptos y definiciones referentes a la Resolución de la Ambigüedad Semántica de las Palabras y métodos de clasificación en general. Luego, se fundamentan formalismos, además de revisiones bibliográficas que demuestra el desarrollado de WSD, así como la actualidad de la investigación. En el capítulo, se identifican, fortalezas y debilidades presentes en propuestas revisadas, con tal de definir las acciones a considerar. Se analizan diferentes bases de conocimiento las cuales son muy utilizadas por la comunidad científica, y finalmente se describen sistemas de competición internacional, para medir la efectividad de la tarea de WSD.

En el Capítulo 3, se abordan primeramente las descripciones del desarrollo de un recurso semántico integrador, capaz de asociar diferentes bases de conocimiento. Esto posibilita, la visualización del entorno de las palabras y sus sentidos desde perspectivas muy distintas. Para ello, se describen dos fases de creación, en la primera se integran cuatro recursos semánticos y en la segunda uno semántico y otro de polaridades-semánticas. Como resultado, se documenta el grado de fiabilidad de la herramienta final a partir de evaluaciones.

En el Capítulo 4 se describen varias propuestas de Resolución de Ambigüedad Semántica, que se consideran no supervisadas y basadas en conocimiento. Para ello, se dividen dos grupos, las Aproximaciones Basadas en Árboles Semánticos, y Basadas en Grafos de Conocimiento. Dentro del primer grupo, básicamente se desea resolver la ambigüedad semántica de las palabras con la extracción de los textos, árboles conceptuales capaces de situar la frase en un determinado contexto. Dentro del segundo grupo, se introduce por primera vez la visión del

modelo de *N-Cliques* en WSD, además de insertar una variante superior en exactitud a la propuesta *Personalizing PageRank* (Agirre and Soroa, 2009). Finalmente en este capítulo, se aplican evaluaciones comparativas, con el fin de conocer en qué medida se han superado o reducido los resultados alcanzados, e incluso considerando reportes emitidos por otros investigadores del área.

En el Capítulo 5 se aplica el Análisis Semántico Multidimensional en el área de la Computación Afectiva en particular en la Minería de Opiniones, donde primeramente se analizan los temas asociados y seguido se plantean las propuestas referentes a resolver las tareas de detección de frases con contenido de opinión, de clasificación de las polaridades de las opiniones y de clasificación de un texto respecto a una pregunta de opinión. Por último se realizan las evaluaciones correspondientes.

Seguido del Capítulo 5, se emiten las conclusiones generales donde se recogen las características relevantes identificadas en la Tesis, además de valorar cada propuesta defendida. Se dan respuesta a los objetivos planteados y se incluyen trabajos futuros que han surgido luego la expansión de la visión semántica multidimensional. Finalmente, dentro de este capítulo se ilustran las producciones científicas referentes a las propuestas abordadas durante todo el trabajo de investigación.

2. ESTADO DEL ARTE

El objetivo de este capítulo es proporcionar la información necesaria para analizar la evolución de la tarea de WSD y de la semántica en general. Para ello se muestran formalismos y conceptos, así como una profunda revisión bibliográfica que demuestra la trayectoria y la actualidad de la investigación. Como parte de la revisión, se agrupan diferentes propuestas de WSD estudiando sus diferentes aproximaciones, con el objetivo de ubicar de esta forma las propuestas defendidas en esta Tesis. Además se van a presentar diferentes recursos semánticos que forman parte de la base de conocimiento de diferentes sistemas de PLN. Y por último y no menos importante se describen las competiciones de evaluación de sistemas que miden la efectividad de la tarea de WSD, demostrando que es un campo que actualmente todavía requiere la atención de la comunidad científica.

2.1. INTRODUCCIÓN

La Desambiguación del Sentido de las Palabras (WSD) es la capacidad de determinar computacionalmente qué sentido es el más adecuado para una palabra por su uso en un contexto particular. Los métodos de WSD consisten en establecer a partir de las definiciones de una palabra, cuál es la definición más adecuada relacionada con el contenido. Usando diferentes técnicas, se identifican una serie de características comunes entre el contexto y las definiciones de la palabra, para más tarde asignar la definición más probable como sentido correcto.

Se puede definir *L* como el texto (bolsa de palabras o secuencia de palabras implicadas en una frase) a analizar, con la siguiente estructura $\{w^1, w^2, w^3, \dots, w^n\}$. Se define formalmente que la tarea WSD otorga el sentido apropiado a cada palabra *w* de *L* o a determinadas palabras del conjunto. Consiste entonces en el alineamiento *A* entre palabras y sentidos. Cada *w* tiene asociado a al menos un sentido o significado *sw* pertenecientes a un diccionario *D*. De modo general la tarea consiste en obtener de la bolsa de sentidos vinculados a una palabra w^k, los más apropiados según el contexto y alinearlos a w^k. Formalmente se representa la alineación *A* de la *k*-ésima palabra w^k como $A(k) \subseteq Sentidos(D, w^k)$, donde $Sentidos(D, w^k)$ es el conjunto de sentidos $\{sw_1^k, sw_2^k, sw_2^k, \dots, sw_n^k\}$ anotados en un diccionario *D* para la *k*-ésima palabra w^k. *A*(*k*) representa el subconjunto de sentidos apropiados al contexto L.

WSD también se ha considerado como una tarea de clasificación, donde los sentidos de las palabras son vistos como clases usando métodos de clasificación automática para asignar cada ocurrencia de las palabras a una o más clases según las evidencias de una fuente externa de conocimiento (Navigli, 2009). Otras tareas estudiadas en esta área de resolución de ambigüedad léxica son, el etiquetado de una parte de discurso (asignación de partes del discurso de palabras objetivo en el contexto); la resolución de la entidad (clasificación de las partidas objetivo textual en categorías predefinidas); la categorización de texto (asignación de etiquetas predefinidas para textos de destino); etc. Una diferencia importante entre estas tareas y WSD es que estas utilizan un único conjunto predefinido de clases, mientras que en WSD el conjunto de clases normalmente cambia en función de la palabra a ser clasificada. Aunque existe la posibilidad de métodos de WSD discriminativos, que no trabajan sobre un conjunto predefinido de sentidos.

Se pueden distinguir dos tipos generales de WSD:

- ***Lexical sample*** (WSD dirigido), donde el objetivo es eliminar la ambigüedad de un conjunto limitado de palabras (comúnmente se selecciona una palabra ambigua por frase). Por lo general, los sistemas supervisados son empleados en este entorno, pues pueden ser entrenados al utilizar un número de casos etiquetados manualmente (conjunto de entrenamiento) y luego se aplican para clasificar un conjunto de ejemplos no etiquetados (de prueba).

- ***All-words*** WSD (todas las palabras en WSD), donde el objetivo es desambiguar todas las palabras de clase abierta en un texto (es decir, sustantivos, verbos, adjetivos y adverbios). Esta tarea requiere de gran cobertura de los sistemas. En consecuencia, los sistemas puramente supervisados pueden sufrir el problema de la escasez de datos, ya que es poco probable que un conjunto de entrenamiento de tamaño adecuado esté disponible. Por otro lado, otros enfoques, tales como los sistemas basados en el conocimiento se basan en los recursos de conocimiento (ej. fuentes de datos como diccionarios, tesauros, etc.) de cobertura completa, donde debe estar asegurada su disponibilidad.

En ambos tipos de tareas se pueden aplicar los sistemas supervisados y no supervisados, aunque los mejores resultados se han obtenido por sistemas supervisados (véase la sección 2.6). Todo sistema de WSD cuenta con los siguientes elementos:

- Proceso de selección de sentidos correcto
- Recursos a tener en cuenta en el análisis de WSD:
    - Contexto de la palabra
    - Fuente de conocimiento externa (ej. recursos léxicos, ontologías, etc.)

2.2. PROCESO DE SELECCIÓN DE SENTIDOS CORRECTOS

Para la selección del sentido correcto de las palabras, primeramente se debe elegir el inventario de sentidos (ej. diccionarios o recursos léxicos en general). En estos inventarios es donde se encuentran las clases a ser asignadas como etiquetas correctas en un proceso de WSD. Para establecer comparaciones entre sistemas de WSD con vista a medir exactitudes y confrontarlas, es necesario establecer diccionarios que puedan ser leídos por los sistemas que contengan inventarios comunes de sentidos. Las clases implicadas deberán estar estructuradas con información semántica, para ser capaces de discernir entre clases representadas por etiquetas similares (ej. palabras polisémicas).

En la obtención del **inventario de sentidos** pertenecientes a un diccionario, se han puesto en vigor varias aproximaciones que estandarizan y definen reglas de procesos de obtención de sentidos. Sin embargo, en la creación del inventario de sentidos influyen varios factores. Como por ejemplo, los diferentes testimonios que sirven como base en la identificación de sentidos, además decidir si el inventario tendrá un nivel de granularidad mayor o menor (este detalle es sumamente importante porque aumenta o disminuye la complejidad de la tarea (ej. véase la reflexión de la sección 2.6.1.4)), qué orden y organización tendrá el inventario, y otras peculiaridades.

Al tener en cuenta estos elementos, se han desarrollado propuestas que se conocen como **aproximaciones generativas**, surgidas para construir inventarios de sentidos. Estas son capaces de generar nuevas etiquetas mediante el proceso de análisis semántico. Es decir, los inventarios son dinámicos o variables según transcurre el proceso. Los sentidos se interpretan como grupos (o *clusters*) de contextos similares de la palabra ambigua. Por ese motivo, los contextos, y los sentidos se pueden representar en **Espacio de la Palabra** y el espacio en el que la proximidad corresponde a la similitud semántica. La similitud en el espacio de palabra se basa en el segundo orden de co-ocurrencia (contextos de la palabra ambigua que se asignan al mismo grupo de sentidos). En las aproximaciones generativas los sentidos pueden ser expresados en términos de "*qualia roles"* (estructura de características semánticas acerca de una entidad) (Hinrich, 1998). La instanciación de una combinación de esta estructura (integrada por roles) permite la creación de un nuevo sentido. Otros enfoques que apuntan a distinciones de sentido más difusas incluyen métodos para la inducción de sentidos. Otros enfoques que apuntan a distinciones sentido más difuso incluyen métodos para la inducción de sentido. A continuación se formaliza la asociación discreta de las distinciones de sentido con palabras codificadas en un diccionario D:

$$Senses_D : Le \times POS \rightarrow 2^C \quad (1)$$

Donde Le representa un lexicón [12], que es el conjunto de palabras codificadas en el diccionario, $POS = \{n, a, v, r\}$ es el conjunto de categorías gramaticales a las que puede pertenecer una palabra (sustantivos, adjetivos, verbos y adverbios respectivamente), y C es el conjunto completo de etiquetas conceptuales en el diccionario D donde 2^C denota el poder conjunto de estos (Navigli, 2009).

Para hacer más fácil la comprensión de estos formalismos es preciso identificar que una palabra w ya categorizada gramaticalmente se define w_p, siendo p la categoría gramatical perteneciente a esa palabra ($p \in POS$), lo que es igual a $w_p \in Le \times POS$. Si se tiene definido w_p, entonces se puede resumir la función $Senses_D(w, p)$ ahora como $Senses_D(w_p)$, con lo que se obtiene entonces el conjunto de sentidos de una palabra categorizada gramaticalmente como el subconjunto $\{sw_p^k{}_1,\ sw_p^k{}_2,\ \ldots, sw_p^k{}_n\}$ donde el sentido se corresponde con el uso que se le dé a la palabra según el contexto donde se utilice. Bajo esta situación se asume la responsabilidad de las herramientas de etiquetado gramatical (en inglés *POS Tagger systems*).

Se entiende como **palabra monosémica** la palabra w_p que al aplicarle la función $Senses_D(w_p)$, se obtiene un conjunto de un solo sentido (lo que es lo mismo $|Senses_D(w_p)| = 1$). Las **palabras polisémicas** son aquellas que pueden tener más de un significado, y adquieren su semántica concreta en función del contexto en que aparecen. A menudo se confunde este fenómeno con otro similar, la **homonimia**; pero de origen totalmente diferente. Dos palabras son homónimas cuando se escriben del mismo modo (**homógrafas**) o suenan de la misma manera (**homófonas**), pero tienen sentidos totalmente diferentes. La homonimia, se refiere a dos o más palabras totalmente diferentes, que provienen de orígenes distintos y poseen significados distintos. En el caso de la polisemia, es una única palabra, con un único origen, la que posee varios significados. Con los siguientes ejemplos quedará más clara esta diferencia:

- Homonimia:
  - dos palabras iguales con dos orígenes y significados diferentes

  Homógrafos:
    - vino (verbo venir, del latín *venit*)
    - vino (bebida, del latín *vinum*)

  Homófonos:
    - tubo (cañería, del latín *tubus*)
    - tuvo (verbo tener, del latín *tenuit*)
- **Polisemia**: una palabra con varios significados: clave:
  - La clave del problema (solución)
  - La clave de la caja fuerte (combinación)

Este último fenómeno (la polisemia), genera la ambigüedad semántica, la cual se trata de resolver mediante métodos de WSD.

[12] Diccionario en el que se registran las palabras que conoce un hablante. Este diccionario especifica los rasgos característicos de las piezas léxicas.

2.3. CONTEXTO DE LA PALABRA

Siendo el contexto el lugar donde se localiza una palabra y además este se encuentra representado por texto no estructurado, se hace necesario aplicarle procesos previos para su futura utilización. Estos procesos se fundamentan en que la información del contexto donde se halla la palabra proporciona información muy útil para ayudar a resolver la ambigüedad (Warren, 1955). Para ello se aplican los pasos siguientes:

- Tokenización (se aplica para la obtención de palabras, en inglés ***tokenization***)
- Asignación de categorías gramaticales (usualmente se conoce en inglés como ***part-of-speech tagging***)
- Lematización (reducción morfológica de la palabra hasta obtener su forma canónica, en inglés ***lemmatization***)
- División del texto para obtener las relaciones directas entre las partes y se conoce en inglés como ***chunking*** (ej. artículo-sustantivo [el jinete])
- Identificación de la estructura sintáctica de la frase (conocido como ***parsing***).

Se puede ver que los tres primeros pasos forman parte del módulo de análisis léxico de los sistemas de PLN descritos en la sección 1.1, luego los siguientes pasos forman parte del módulo sintáctico. En la Figura 2 se muestra un ejemplo del proceso tratamiento que se le aplica al corpus.

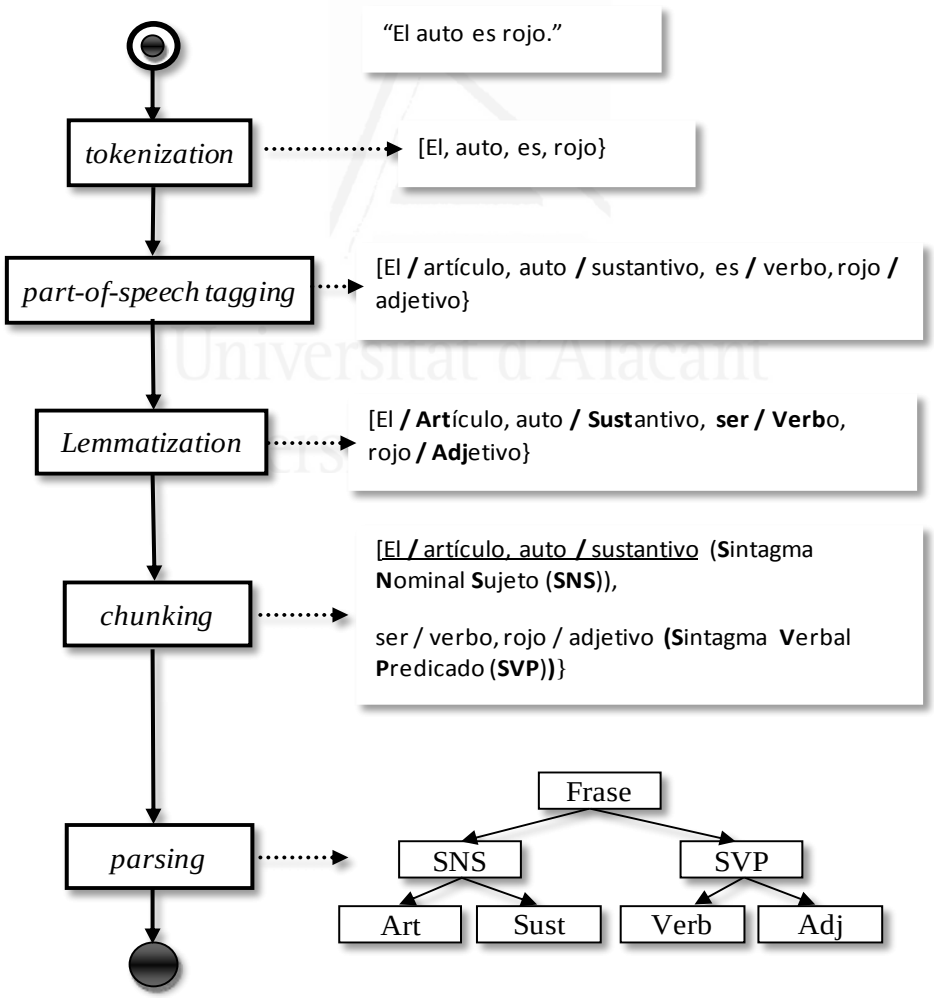


Figura 2. Pasos de pre-procesamiento del contexto.

Gracias a la ejecución de este pre-procesamiento es posible entonces obtener valiosas informaciones del contexto:

- **Características locales** (ej. etiquetas del lenguaje, formas de las palabras, las posiciones con respecto a la palabra objetivo, etc.)
- **Características del tópico** (contrario a las locales definen características generales del tópico del texto o discurso (ej. una ventana de palabras, una oración, una frase, a un párrafo, etc.), usualmente como bolsa de palabras)
- **Características sintácticas** (encuentran relaciones claves entre las palabras del contexto, es importante remarcar que algunas palabras pudieran encontrarse fuera de contexto)
- **Características semánticas** (identifica informaciones de significado de palabras (ej. sentido de las palabras), contexto asociado a dominios, etc.)

Teniendo en cuenta las características que se pueden obtener es posible construir el ejemplo de la Tabla 1 y Tabla 2 a partir de la oración: El jugador principal logró un gol fantástico. Se analiza con respecto a la palabra jugador.

w1-	w2-	w1+	w2+	Dominio de Oración
artículo	-	adjetivo	verbo	Deporte

Tabla 1. Características con dos elementos como ventana de palabras.

Como se observa en la Tabla 1 se ha realizado un análisis de los elementos adyacentes de la palabra analizada a diferentes distancias según sus posiciones en la oración. Esto hace que el contexto analizado varíe. Esta acción se conoce comúnmente como ventana de palabras y en ella intervienen el número de gramas seleccionados. Los n-gramas son una secuencia de n palabras incluyendo la palabra en cuestión (ej. unigrama ($n = 1$), bigrama ($n = 2$), trigrama ($n = 3$),...hasta incluir toda la frase, incluso todo el párrafo). Entonces se dice que la ventana de palabras se construye con los gramas recién mencionados o de manera siguiente, con $\pm n$ es $(2n + 1)$, esto se aplica de este modo porque se induce que la palabra en cuestión tiene que estar incluida y se pueden ofrecer todas las combinaciones con $(2n + 1)$-gramas. En la Tabla 2 se ilustra un ejemplo.

Gramas	Conjuntos de palabras
unigrama	jugador
bigrama	el jugador
	jugador principal
trigrama	__ el jugador
	El jugador principal
	jugador principal logró
Ventana (n = 2)	jugador principal logró un gol
(2n + 1)-gramas	El jugador principal logró un
oración	El jugador principal logró un gol fantástico.
párrafo	El jugador principal logró un gol fantástico. Le darán la zapatilla de oro.

Tabla 2. Ventana de palabras respecto a jugador.

El uso de los gramas no es la única vía para conseguir el contexto, se puede considerar también como contexto la estructura sintáctica de la oración. Otra variante es la generación de árboles semánticos (aquí se incluye la búsqueda de ancestros semánticos) (Agirre 1996, Gutiérrez et al., 2011b, Mihalcea and Moldovan, 1998), la red de coocurrencia presentada en documentos (Widdows and Dorow, 2002, Veronis, 2004), la construcción de grafos semánticos (Agirre and Soroa, 2009, Sinha and Mihalcea, 2007, Laparra et al., 2010), etc.

En la aplicación de métodos supervisados de WSD generalmente es más apropiado el uso de vectores de características (Navigli, 2009). En contraste, las representaciones estructuradas son más útiles en los métodos de supervisión basada en el conocimiento, ya que pueden aprovechar plenamente las relaciones léxicas y semánticas entre los conceptos codificados en las redes semánticas y léxicos computacionales. Es correcto señalar que la elección del tamaño adecuado

del contexto es un factor importante en el desarrollo de un algoritmo de WSD, ya que afecta el desempeño de la desambiguación (véase, por ejemplo, (Cuadros and Rigau, 2006)).

2.4. FUENTES DE CONOCIMIENTO EXTERNAS

La creación de fuentes de conocimiento externa se centra en el diseño de formalismos cognitivos y computacionalmente apropiados para expresar el conocimiento en un área particular. Se ha representado en los sistemas de información de diversas maneras, tal y como se describe en (Samper, 2005). Se dice entonces que tienen puntos de convergencia en múltiples áreas tecnológicas, por ejemplo:

- En las bases de datos se han utilizado diagramas entidad-relación para definir los conceptos y sus relaciones en un determinado universo.
- En programación se han utilizado gramáticas y estructuras de datos como clases y objetos.
- En Ingeniería del Software se ha propuesto el uso de lenguajes de modelado como UML, donde también es posible definir clases y sus relaciones.
- En algunas de las formas que se han utilizado para representar el conocimiento en Inteligencia Artificial, son la lógica matemática y las estructuras de datos. En la lógica matemática destacan los enfoques basados en lógica y los basados en reglas.

En relación con los sistemas de representación del conocimiento basados en estructuras de datos, se destacan los siguientes:

- Redes semánticas
- Marcos (*frames*)
- Redes de herencia estructurales
- Sistemas terminológicos o Descripciones Lógicas
- Grafos, Redes de Petri, Mapas Tópico

En los últimos años las Ontologías son vistas como una forma más avanzada de representación del conocimiento. También es muy utilizada la combinación de estas formas, por ejemplo, *frames* con reglas, objetos y relaciones, etc. Con lo cual, los componentes fundamentales externos para WSD están constituidos por fuentes de conocimiento muy complejas o, pueden presentarse con diferentes naturalezas. Pueden ser desde la información presente en una base de datos relacional, o bases de conocimiento como las recientemente mencionadas, o corpus de textos ya sea sin etiquetar o anotados con sentidos de las palabras, de diccionarios digitales, tesauros, glosarios, ontologías, etc. Todas ellas proveen información vital en la ardua tarea de relacionar las palabras con las múltiples definiciones con que cuentan. A continuación se presenta una breve demostración de estos exponentes.

- **Recursos estructurados**
  - **Tesauros** (en inglés *Thesauri*): provee relaciones entre palabras (sinónimos, antónimos, etc.). Entre los tesauros más utilizados se encuentra "*Roget's International Thesaurus of English Words and Phrases*" [13], con 85.000 hipervínculos de referencias cruzadas, cumpliendo así su objetivo de "cada palabra está relacionada con sus vecinos y cada parte con el todo." Además, más de 2.900 proverbios y citas de autores clásicos y modernos ilustran las más de 1000 entradas.

[13] http://www.bartleby.com/110/

- **Diccionarios digitales** (en inglés *Machine-readable dictionaries* (MRDs)): Como su nombre lo indica son diccionarios disponibles en formato electrónico. Los más utilizados por ejemplo son el "*Collins English Dictionary"*[14], de *Oxford Advanced Learner's Dictionary of Current English*, de la Universidad de Oxford. También el "*Longman Dictionary of Contemporary English"* [15] (LDOCE). Luego se inició la aparición de WordNet[16] (creado en Princenton) (Miller et al., 1990) el que superó en uso al anterior diccionario mencionado, este constantemente se enriquece y cuenta con una estructura de red semántica de conceptos. Por esta razón en varias literaturas se le puede encontrar con la definición ontología aunque no es del todo cierto (véase la definición de ontología a continuación). Y por último y no menos importante Wikipedia[17] que figura como una enciclopedia ha aumentado sus utilidades hasta el punto de convertirse en uno de los recursos electrónicos más difundidos en la actualidad (Santamaría, 2010).
- **Ontologías** (en inglés *Ontologies*): Es una conceptualización de un dominio con la finalidad de compartir cierta información entre diferentes agentes. Se define un vocabulario común con una serie de conceptos y relaciones que se emplea para la comprensión de un área o dominio (Hovy, 2003, Montoyo Guijarro, 2008). El objetivo de las ontologías es facilitar las descripciones, las búsquedas semánticas y el razonamiento (Hovy, 2003). Comenta (Gruber, 1993) *"An ontology is a explicit specification of a conceptualization"* donde su traducción al español significa que una ontología es una especificación explícita de una conceptualización.

- **Recursos no estructurados**
  - **Corpus** (en inglés *Corpora*): Colecciones de textos usados para modelar el aprendizaje del lenguaje, se pueden presentar con sentidos anotados o sin anotar. Utilizados indistintamente por sistemas WSD supervisados y no supervisados.
    - **Corpus no etiquetado** (en inglés *Raw corpora*): Un ejemplo de corpus no etiquetado o corpus crudo es el *Brown Corpus*[18], con una colección de millones de palabras de los textos publicados en los Estados Unidos en 1961. Otro corpus que se puede mencionar es el *British National Corpus*[19] (BNC), de 100 millones de palabras de recogida de muestras habladas y escritas del idioma inglés (a menudo utilizado para recoger las frecuencias de palabras e identificar las relaciones gramaticales entre las palabras), y otros.
    - **Corpus con sentidos etiquetados** (en inglés *Sense-Annotated Corpora*): Textos semánticamente anotados basándose en inventarios predefinidos. El corpus de SemCor [20] es el corpus más usado por sistemas y competiciones de PLN. Consiste en la anotación semántica del Corpus del Brown con versiones de WordNet, la cual incluye 234136 sentidos anotados en 352 textos. Otro exponente, es el corpus MultiSemCor [21], anotado paralelamente en idioma inglés e italiano. Cuenta con 116 textos de inglés con sus correspondientes traducciones

[14] http://www.collinslanguage.com/
[15] http://www.ldoceonline.com/
[16] http://www.cogsci.princeton.edu/~wn/
[17] http://es.wikipedia.org/wiki/Wikipedia:Portada
[18] http://www.archive.org/details/BrownCorpus
[19] http://www.natcorp.ox.ac.uk/
[20] http://www.cse.unt.edu/~rada/downloads.html
[21] http://universal.elra.info/product_info.php?cPath=42_43&products_id=1627

al italiano, para un total de alrededor de 500.000 *tokens*[22]. Este se basa en el corpus SemCor, los textos están alineados a nivel de palabra y contienen anotaciones semánticas. Otro es DSO (*Defence Science Organisation*) *Corpus of Sense-Tagged English*[23], que incluye 121 sustantivos y 70 verbos los cuales son los más frecuentes que ocurren en palabras ambiguas del idioma inglés; estos resultados son provienen del análisis de alrededor de 192,800 frases del corpus del Brown y del *Wall Street Journal*[24]. El *Open Mind Word Expert*[25] anotado con 288 sustantivos, en posteriores versiones se adicionaron los verbos y adjetivos, este corpus fue utilizado en la competición Senseval-3[26]. Otro de los corpus más usados tanto en las competiciones que le dan nombre como en comparativas de investigaciones científicas son los corpus de las ediciones de Senseval[27] (Yarowsky et al., 2001, Mihalcea et al., 2004, Snyder and Palmer, 2004, Agirre et al., 2007, Navigli et al., 2007, Pradhan et al., 2007). Todos los corpus que han sido mencionados se han anotado por diferentes versiones de WordNet excepto el de la competición de Senseval-1. Aunque existen algunos más, estos son los que ofrecen una muestra representativa de las fuentes anotadas que se utilizan en aplicaciones de WSD. Algunos de esos corpus son evaluados y analizados más adelante en esta Tesis.

- **Recursos de Colocación** (en inglés *Collocation resources*), este tipo de recursos registran la tendencia que tienen las palabras para aparecer con regularidad cercanas a otras, un ejemplo de este tipo de recursos puede ser *Word Sketch Engine*[28], *Just The Word*[29], *The British National Corpus collocations*[30], etc. Este tipo de recurso se beneficia con el suministro de enormes cantidades de información, y a su vez los sistemas de WSD.
- **Otros recursos**, uno de los recursos más efectivos en WSD, aunque no indica propiamente semántica, pero suelen ser muy efectivos son la listas de frecuencias de palabras (Mc.Carthy *et al.*, 2004) (realizan un conteo de cuantas apariciones presentan las palabras en multitudes de documentos), las listas de *stopwords* (listas de palabras no discriminada sin contenido (como: un, una, el, la,...etc.)), etiquetas de dominio[31], etcétera.

Es importante profundizar un poco más en las ontologías y redes semánticas, por ser de los exponentes más ricos en información semántica en la actualidad. Según (Hovy, 2003) de ellas se pueden obtener ciertas bondades como son:

- Abarcar terminologías específicas del dominio, además de con representaciones detalladas que distinguen sus datos, lo cual permite ser consultada por expertos y sistemas informáticos.

[22] Componente léxico es una cadena de caracteres que tiene un significado coherente en cierto lenguaje de informático.
[23] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12
[24] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43
[25] http://www.cse.unt.edu/~rada/downloads.html#omwe
[26] http://www.senseval.org/senseval3/
[27] http://www.senseval.org/senseval3/
[28] http://www.sketchengine.co.uk/
[29] http://www.just-the-word.com/
[30] http://bncweb.info/
[31] http://wndomains.fbk.eu/index.html

- Incluir terminologías que posibiliten la localización rápida de la información (Tello, 2001).
- Ayudar en la inferencia automática para que los sistemas puedan sugerir a los usuarios la información más cercana a la búsqueda y con esto proponer un conjunto de datos relacionados localizados en el mismo dominio.
- Facilitar la incorporación semiautomática de información de nuevos dominios.

En el ámbito de la Inteligencia Artificial, un sistema inteligente solamente conoce lo que se puede representar en un lenguaje determinado. Por este motivo, es razonable utilizar el término **ontología** para designar **lo que el sistema conoce**, donde el conjunto de objetos que se representan se denominan el **universo del discurso**. Este conjunto de objetos y las relaciones entre estos, son plasmados mediante un vocabulario de representación con el que un programa describe el conocimiento.

Se puede describir la ontología como un conjunto de términos representacionales. En ella, las definiciones relacionan nombres de entidades con el universo de discurso (clases, relaciones, funciones y otros objetos) usando texto que puede ser leído y que describe lo que significan los nombres y los axiomas formales que restringen la interpretación.

Los términos que utilizan las ontologías son los siguientes:

- **Conceptos**: colecciones de objetos del dominio.
- **Relaciones**: representan interacciones entre conceptos del dominio (subclase-de, parte-de,...).
- **Funciones**: tipo especial de relación (tiempo-duración, días-que-faltan,...).
- **Instancias**: representan elementos determinados de una clase.
- **Axiomas**: teoremas que se declaran sobre relaciones que deben cumplirse entre elementos de la ontología.

Se clasifican de diferentes formas (Montoyo Guijarro, 2008):

- **De contenido**
  - Construidas para reutilizar su conocimiento.
- **De indexación**
  - Permiten la recuperación de casos cuando los agentes comparten conocimientos a través de BD.
- **De comunicación**
  - Usadas por agentes para obtener respuestas a preguntas concretas.
- **Meta-ontologías**
  - Utilizadas para representar ontologías.
- **De representación**
  - Proporcionan el vocabulario necesario para modelar otras ontologías (Moldovan and Rus, 2001).
- **Genéricas**
  - Proporcionan términos genéricos reutilizables en diferentes dominios.
- **De dominio**
  - Expresan conceptualizaciones que son específicas para dominio particulares (Sara and Daniele, 2009).
- **De aplicación**
  - Contienen todas las definiciones que son necesarias para modelar los conocimientos requeridos por una aplicación particular.

2.4.1. LENGUAJES DE REPRESENTACIÓN DEL CONOCIMIENTO

Las ontologías por lo general se almacenan físicamente mediante su escritura en los diferentes lenguajes de representación que se enumerarán a continuación:

- Lógicos
  - KIF (*Knowledge Interchange Format*)[32]
  - CycL / OpenCycL[33]
  - Loom (PowerLoom[34])
  - Frame-Logic (F-Logic[35])
- De marcado
  - SHOE[36]
  - XOL (*Ontology Exchange Language*[37])
  - OML / CKML (*Ontology Markup Language*[38])
  - RDF / RDFS (*Resource Description Framework Schema Language*[39])
  - OIL / DAML+OIL (*Ontology Interchange Language*[40])
  - OWL (*Ontology Web Language*[41])

De todos estos los más difundidos son los tres últimos, debido a la riqueza de conocimiento descrita por formalismos presentes en las relaciones internas entre elementos. Es importante destacar, que entre los diversos tipos de lenguajes de representación aquí señalados, estos difieren por su naturaleza (lógicos y de marcado). El más difundido en la actualidad es OWL, este se ha concebido con el objetivo de compartir conocimiento a través de la Web, sustentando entonces lo que hoy se conoce como la Web semántica. Este lenguaje se soporta sobre las bases de RDF (en muchas ocasiones se encuentran descritos OWL/RDF), al reutilizar sus esquemas y extenderlos de modo que su riqueza semántica sea cada vez más expresiva. De OWL se han desarrollado tres categorías, OWL *Lite*, OWL DL y OWL *Full*, con el objetivo de que en cada una se expresen diferenciados tipos de elementos (ej. axiomas de Clase, combinaciones de expresiones *Boolean*, arbitrariedades de cardinalidad y filtrados de información) que haga que cada categoría sea más compleja que la otra respectivamente. Como aspecto fundamental que se ha de señalar en OWL, se tiene que cada definición de Clase o propiedad asociada la creación de instancias, siempre tendrá asociado una URI [42] (*Uniform Resource Identifier*) como identificador único y distinguible en la Web. De este modo se logran estandarizar y reutilizar las ontologías.

Para conocer un poco de RDF (*Resource Description Framework*) se puede decir que este es un fundamento para el procesamiento de metadatos. Proporciona interoperabilidad entre aplicaciones donde se intercambian información en la Web. Basado en XML, este extiende todavía más sus especificaciones para lograr describir recursos de cualquier tipo (incluyendo recursos XML y no-XML). RDF puede utilizarse en diferentes áreas de aplicación (ej. para mejorar las capacidades de motores de búsqueda de recuperación de recursos, en catalogación y

[32] http://logic.stanford.edu/kif/
[33] http://www.opencyc.org/
[34] http://www.isi.edu/isd/LOOM/PowerLoom/
[35] http://www.cs.umbc.edu/courses/771/papers/flogic.pdf
[36] http://www.cs.umd.edu/projects/plus/SHOE/
[37] http://www.ai.sri.com/pkarp/xol/
[38] http://www.ontologos.org/OML/OML%200.3.htm
[39] http://www.w3.org/standards/techs/rdf#w3c_all
[40] http://www.w3.org/TR/2001/NOTE-daml+oil-reference-20011218
[41] http://www.w3.org/standards/techs/owl#w3c_all
[42] http://www.w3.org/TR/uri-clarification/

relacionamiento de contenidos accesibles en la Web o biblioteca digital, para compartir información entre agentes inteligentes, para la descripción de los derechos de propiedad intelectual de las páginas Web, para expresar políticas de privacidad de un sitio Web, y otras). Se considera entonces que junto con las firmas digitales, RDF será la clave para construir el "Web de confianza" (Klyne and J.Carroll, 2006) del comercio electrónico, la colaboración y otras aplicaciones. Como se puede observar, RDF tiene un nivel tal de desarrollo, que hace muy común que en la literatura aparezca RDF/OWL como un mismo lenguaje, pero en realidad OWL es más descriptivo.

Para la creación de Ontologías ya sea utilizando lenguajes lógicos como de marcado, se han creado varias herramientas de desarrollo de ontologías. En (Corcho *et al.*, 2002) se describen una conjunto que se mencionan a continuación:

- Apollo
- LinkFactory ®
- OILEd
- OntoEdit
- Ontolingua Server
- OntoSaurus
- OpenKnoME
- Protégé
- SymOntoX
- WebODE
- WebOnto

Uno de los recursos semánticos más utilizados en PLN es WordNet (WN). Este se ha escrito con diferentes estructuras, ya sea estandarizada (con lenguajes ontológicos) o no.

2.4.2. WORDNET

Es una base de datos léxica para el idioma inglés, también considerada como una ontología, fue creada por la Universidad de Princenton[43]. WordNet (Miller et al., 1990) representa una red semántica conceptual y estructurada, en ella se tienen en cuenta sustantivos, verbos, adjetivos y adverbios. La unidad básica es el *synset* (*synonym sets* o conjuntos de sinónimos), este representa un concepto de forma léxica (Ševčenko, 2003). Un *synset* se identifica por número único de ocho dígitos llamado *offset* (número que coincide con su posición el fichero). Dentro de la base de datos, cada *synset* representa un significado 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 interrelaciones entre significados de palabras.

Las relaciones semánticas que se establecen de forma general entre *synsets* son las siguientes:

- Sinonimia
- Antonimia
- Hiponimia / Hiperonimia
- Meronimia / Holonimia
- Implicación y Causa, y otras

[43] http://wordnet.princeton.edu/wordnet/

Se indica *synset* al conjunto de palabras que son sinónimos en un mismo sentido. Además de distinguir mediante *synsets* los significados de cada término, WN establece una relación de orden entre los diferentes sentidos de las palabras, de acuerdo a su frecuencia de aparición (ej. en el corpus de SemCor). De esta forma, para ***image*** en la versión 2.0 de WordNet existen ocho significados mostrados en la Tabla 3. Como se observa para una sola palabra se establecen varios significados y de cada uno se conoce el conjunto de sinónimos ordenados en cada caso por la frecuencia de uso según el sentido, y además una frase que lo describe (glosa).

***Synsets* y Categorías gramaticales**
<<<<<< Sustantivo >>>>>>
00039630 -- -- effigy -- -- image -- -- simulacrum -- a representation of a person (especially in the form of sculpture); "the coin bears an effigy of Lincoln"; "the emperor's tomb had his image carved in stone"
<<<<<< Sustantivo >>>>>>
00043454 -- -- picture -- -- image -- -- icon -- -- ikon -- a visual representation (of an object or scene or person or abstraction) produced on a surface; "they showed us the pictures of their wedding"; "a movie is a series of images projected so rapidly that the eye integrates them"
<<<<<< Sustantivo >>>>>>
00047709 -- -- persona -- -- image -- (Jungian psychology) a personal facade that one presents to the world; "a public image is as fragile as Humpty Dumpty"
<<<<<< Sustantivo >>>>>>
00053779 -- -- image -- -- mental_image -- an iconic mental representation; "her imagination forced images upon her too awful to contemplate"
<<<<<< Sustantivo >>>>>>
00053832 -- -- prototype -- -- paradigm -- -- epitome -- -- image -- a standard or typical example; "he is the prototype of good breeding"; "he provided America with an image of the good father"
<<<<<< Sustantivo >>>>>>
00059547 -- -- trope -- -- figure_of_speech -- -- figure -- -- image -- language used in a figurative or nonliteral sense
<<<<<< Sustantivo >>>>>>
00074537 -- -- double -- -- image -- -- look-alike -- someone who closely resembles a famous person (especially an actor); "he could be Gingrich's double"; "she's the very image of her mother"
<<<<<< Verbo >>>>>>
00109926 -- -- visualize -- -- visualise -- -- envision -- -- project -- -- fancy -- -- see -- -- figure -- -- picture -- -- image -- imagine; conceive of; see in one's mind; "I can't see him on horseback!"; "I can see what will happen"; "I can see a risk in this strategy"

Tabla 3 Sentidos de ***image***.

Es conveniente destacar que este recurso léxico se ha extendido a diferentes idiomas, se puede encontrar en inglés, español (castellano), holandés, italiano, alemán, francés, checo, estonio, sueco, noruego, danés, griego, portugués, vasco, catalán, rumano, lituano, ruso, búlgaro, esloveno, y otros que están en proceso actualmente[44]. Estas variantes del lenguaje, se han desarrollado inicialmente bajo la tutela de la Universidad de Princeton y posteriormente sustentado por la Asociación de WordNet Global[45]. Recursos como EuroWordNet (Vossen, 1998), MultiWordNet (Pianta *et al.*, 2002), *Multilingual Central Repository* (MCR) (Atserias *et al.*, 2004) y otros, se refieren a WN como núcleo léxico. Esta percepción resulta muy interesante, comprobándose que tomando como núcleo a WN, es posible lograr la integración semántica.

[44] http://www.illc.uva.nl/EuroWordNet/
[45] http://www.globalwordnet.org/

2.4.2.1. RECURSOS SEMÁNTICOS ALINEADOS A WORDNET

Debido a la enorme repercusión que ha tenido el uso de la base de datos léxica de WN en investigaciones de PLN, destacados investigadores se propusieron elaborar diferentes recursos léxicos-conceptuales que tuvieran vínculos internos con los elementos fundamentales de WN (los *synsets*), con el objetivo de generar mayor conocimiento. Algunos de estos recursos se crearon partiendo del mismo WN y otros surgieron a partir de la asociación de etiquetas pre-elaboradas y concebidas estructuralmente. A continuación se describirán algunos de estos recursos.

2.4.2.1.1. WORDNET *DOMAINS*

Es un recurso anotado en idioma inglés, WordNet *Domains* (WND) extiende la información proporcionada por WordNet mediante la inclusión de *Subject Field Codes* (SFC) (Magnini and Cavaglia, 2000), es decir, conjuntos de palabras relevantes para un dominio específico. Por un lado, estas etiquetas conceptuales clarifican a qué contexto se refiere la definición y por otro, permiten la búsqueda rápida del concepto deseado. Por ejemplo, si se busca el significado de disco dentro del contexto de la Informática, simplemente basta con mirar la etiqueta del campo semántico que precede a cada definición hasta dar con la que interesa. Con el fin de incorporar la información de las etiquetas semánticas a WN, se construyó este nuevo recurso. Con esta nueva herramienta se pretende mejorar la distinción de los sentidos en WN, al agrupar en muchos casos distintos sentidos bajo un mismo dominio o categoría semántica. En WND los *synsets* de WN han sido anotados mediante un proceso semi-automático con una o varias etiquetas de dominio, seleccionadas 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 gramaticales.
- 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.
- Dentro de un mismo dominio pueden incluirse *synsets* que pertenecen a diferentes categorías sintácticas.
- Dentro de un mismo dominio pueden aparecer sentidos de palabras pertenecientes a diferentes sub-jerarquías de WN.
- 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***, tiene en WN diez sentidos (Ver Tabla 4). Al observar esta tabla se puede reducir el nivel de polisemia de diez sentidos a cuatro sentidos, agrupando aquellos sentidos que pertenecen a un mismo dominio (Magnini *et al.*, July 2002). En la jerarquía de dominios se encuentran diferentes niveles de especificación. Mientras más se profundiza es mayor el nivel de especialización de los dominios. En la Figura 3 se muestra un pequeño fragmento de la jerarquía de WND tomado de (Luisa Bentivogli, 2005).

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. Sentidos y correspondientes dominios de *man*.

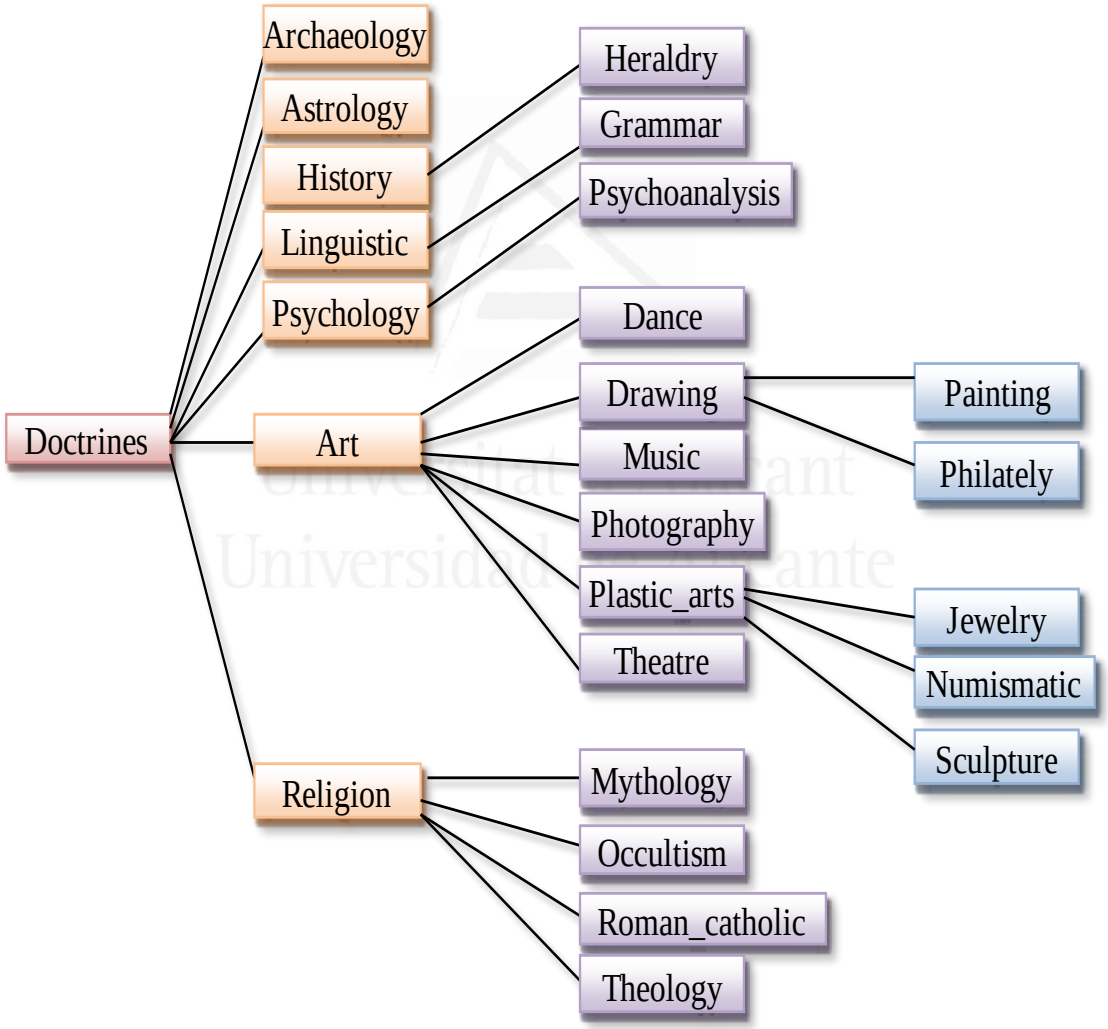


Figura 3. Jerarquía de WordNet *Domains*.

Entre el conjunto de SFC's se aprecia una etiqueta de dominio denominada *Factotum*. Este dominio se ha creado exclusivamente para englobar dos tipos de *synsets*(*Magnini and Cavaglia, 2000*):

- *Synsets* genéricos. Son aquellos que difícilmente se pueden clasificar dentro de algún dominio en particular.
- *Stop senses*. Son aquellos que aparecen frecuentemente en diferentes contextos, tales como números, días de la semana, colores, etc.

2.4.2.1.2. WORDNET *AFFECTS*

WordNet *Affects* (WNA) constituye una extensión de WND (Magnini and Cavaglia, 2000, Sara and Daniele, 2009), se comprende por subconjuntos de conceptos afectivos que agrupan *synsets* que denotan estados emocionales. Las etiquetas fueron anotadas con un proceso similar a WND. Algunos de los conceptos que representa son *moods* (humores), *situations eliciting emotions* (situaciones que afloran emociones) o *emotional responses* (respuestas emocionales).

Este recurso fue extendido con un conjunto de etiquetas adicionales llamadas *emotional categories* (categorías emocionales). Se estructura de forma jerárquica al utilizar como relación entre conceptos afectivos la hiperonimia de WordNet (Valitutti *et al.*, 2004). En una segunda revisión de este recurso se introdujeron algunas modificaciones a la hora de distinguir qué sentidos están más acordes con las etiquetas emocionales y también se incluyeron nuevas como son: *positive*, *negative*, *ambiguous* y *neutral:*

- La primera corresponde a las emociones positivas, por ejemplo incluye *synsets* como*: joy#1* o *enthusiasm#1*.
- La negativa define como su nombre lo indica signos negativos como por ejemplo*: anger#1* o *sadness#1*.
- Los ambiguos representan *synsets* que su estado afectivo depende de la semántica del contexto: *surprise#1*
- Los neutrales representan a los *synsets* que se refieren a estados mentales que no se caracterizan por tomar partido.

Una propiedad importante, es que las etiquetas asocian a los adjetivos y sustantivos que se ven implicados en el uso de estados de emociones. Como el adjetivo modifica el estado del sustantivo, en ocasiones se puede determinar el estado del sustantivo modificado, por ejemplo: *cheerful* (alegre) / *happy boy* (chico feliz)) (Strapparava and Valitutti, 2004). Es decir, con conocer si el adjetivo obedece a un estado emocional se indica cómo se encuentra el sustantivo. A continuación se muestra la Tabla 5 con una lista de etiquetas afectivas a los que les corresponden *synsets*.

Etiquetas afectivas	Ejemplos
emotion	*noun anger#1, verb fear#1*
mood	*noun animosisy#1, adjective amiable#1*
trait	*noun aggressiveness#1, adjective competitive#1*
cognitive state	*noun confusion#2, adjective dazed#2*
physical state	*noun illness#1, adjective all in#1*
hedonic signal	*noun hurt#3, noun suffering#4*
emotion-eliciting situation	*noun awkwardness#3, adjective out of danger#1*
emotional response	*noun cold sweat#1, verb tremble#2*
behavior	*noun offense#1, adjective inhibited#1*
attitude	*noun intolerance#1, noun defensive#1*
sensation	*noun coldness#1, verb feel#3*

Tabla 5. Etiquetas de WordNet *Affects* y correspondientes *synsets.*

2.4.2.1.3. SUMO

SUMO[46] (*Suggested Upper Merged Ontology*) se considera una ontología de nivel superior. Proporciona definiciones para términos de propósito general y puede actuar como base para ontologías de dominios más específicos. Fue creada a partir de la combinación de diferentes contenidos ontológicos en una única estructura cohesiva. Actualmente existen alrededor de 1000 términos y 4000 aserciones (Niles and Pease, 2003).

Los contenidos a partir de los cuales se obtuvo SUMO, proceden de: Ontolingua[47] y las ontologías desarrolladas por ITBM-CNR [48] (*Unrestricted-Time, Representation, Anatomy, Biologic-Functions, and Biologic-Substances*). El lenguaje de representación estándar utilizado es una versión de KIF (*Knowledge Interchange Format*) (Genesereth and Fikes, 1992), llamada SUO-KIF (Pease, 2007).

Para su concepción 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 de Russell y Norvig (Russell and Norvigy, 1994). En el segundo grupo se incluyó el resto. Luego, las dos ontologías de alto nivel se combinaron entonces para obtener una única estructura conceptual. El resto del contenido de las clases de bajo nivel fue añadido tras la combinación. Para comprender la estructura y el contenido de SUMO se pueden extraer los conceptos de más Alto nivel (ver Figura 4).

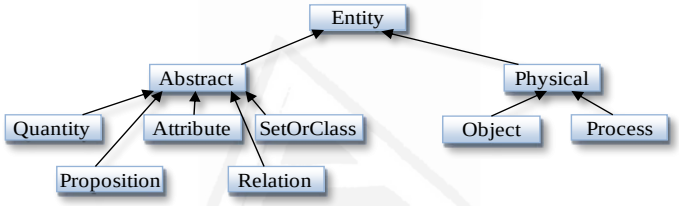


Figura 4. Conceptos de alto nivel de SUMO (Ševčenko, 2003).

Al igual que en la mayoría de las jerarquías el concepto de más alto nivel la categoría ***entity*** engloba a todos las demás y bajo este concepto se encuentran ***physical*** y ***abstract***. Un ejemplo para ***bank#1*** se puede observar en la Figura 5, se ilustra la estructura de SUMO.

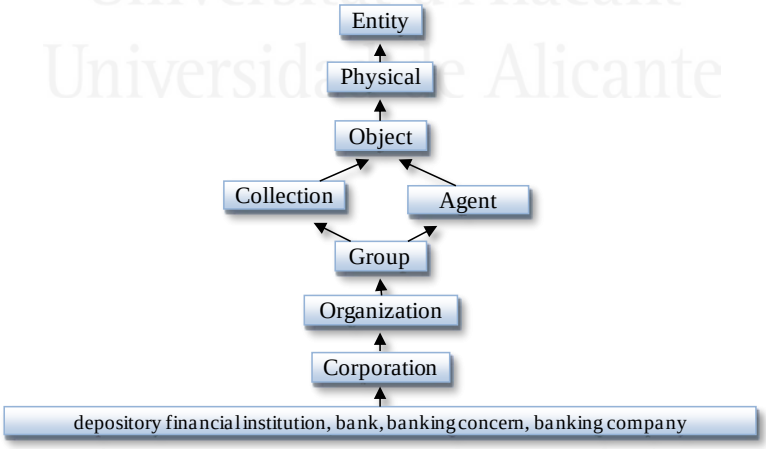


Figura 5. Jerarquía de SUMO para ***bank#1***.

[46] http://suo.ieee.org/SUO/SUMO/index.html
[47] http://www.ksl.stanford.edu/software/ontolingua/
[48] http://www.ontologyportal.org/SUMOhistory/

Existen muchos más recursos que se alinean a WordNet como son *Extended* WordNet (Moldovan and Rus, 2001), FrameNet (Sara and Daniele, 2009) , recursos multilingües y otros. Seguidamente se nombrarán algunos recursos de integración semántica que tienen como núcleo a WordNet.

2.4.2.1.4. CLASES SEMÁNTICAS

Una Clase Semántica es una conceptualización de sentidos que pueden ser creadas manualmente o semiautomáticamente, de diferentes niveles de abstracción, y centradas en dominios variados. Al analizar WN se puede decir que está compuesto por una serie de *synsets* conectados y relacionados entre sí por diversas relaciones semánticas. Cada uno de estos *synsets* representa un concepto y contiene un conjunto de palabras que hacen referencia a ese concepto, y por tanto son sinónimos. Estos *synsets* están clasificados en cuarenta y cinco grupos en función tanto de categorías léxicas (adjetivos, nombres, verbos y adverbios), como de agrupamientos semánticos (persona, fenómeno, sentimiento, lugar, etc.) (Fellbaum, 1998). Existen veintiséis categorías para nombres, quince para verbos, tres para adjetivos y uno para adverbios (Izquierdo, 2010).

Este diseño organizacional se ha hecho para que cada lexicógrafo especialista en un área en particular, pueda obtener una estructura con la cual crear y editar el conjunto de palabras y significados bajo una misma clase semántica. Estas categorías semánticas se consideran también como Clases Semánticas más generales que sus sentidos, donde, varios sentidos de WN se agrupan bajo determinadas clase semántica, además se tiene en presentes los usos más frecuentes y generales de las palabras. Esta conceptualización de sentidos está disponible para cualquier lengua contenida en el recurso EuroWordNet, ya que todos los WNs están enlazados con el WordNet en inglés.

Es importante destacar que, a pesar que en un proceso *top-dwon* (desde arriba hacia abajo) se generó WN, normalmente no es posible distinguir diferencia alguna entre *synsets* que un momento fueron Clases Semánticas y los que no. Entonces para recuperar esta categorización de sentidos, se debe aplicar algún método de generación de Clases Semánticas.

La generación de Clases Semánticas (*Semantic Classes* (SC)) (Izquierdo *et al.*, 2007) se ha orientado hacia el objetivo de reducir la polisemia, y para ello se han desarrollado diferentes técnicas donde se ejecutan agrupaciones de sentidos. En todos los casos se han agrupado sentidos de la misma palabra, reduciéndose la polisemia e incrementándose los resultados de un sistema de desambiguación.

El recurso de SC consiste en un conjunto de Conceptos de Nivel de Base (*Base Level Concepts*) (BLC) obtenidos a partir de WN luego de aplicar un proceso de *botton-up* (desde abajo hacia arriba), con la cadena de las relaciones de hiperonimia (hijo-padre). Para cada *synset* de WN, el método selecciona como su concepto de nivel de base el primer máximo local según el número relativo de las relaciones establecidas para la creación del recurso. Como resultado, las clases semánticas tienen un conjunto de BLC's que son semánticamente vinculadas a varios *synsets*.

El proceso en cuestión sigue un recorrido ascendente a través de la cadena de hiperonimia definida mediante las relaciones correspondientes en WordNet. Para cada *synset*, se selecciona como su BLC apropiado el primer máximo local de acuerdo al número relativo de relaciones. En caso que el *synsets* tenga varios hiperónimos, se selecciona el camino que tenga el máximo local, con el mayor número de relaciones. El proceso termina con un conjunto de conceptos candidatos iníciales, con el conjunto de *synsets* seleccionados como BLC para algún otro *synset*. Entre estos, existen BLC que no representan a un número suficiente de conceptos o *synsets*. Para evitar estos falsos BLC's, se realiza un proceso de filtrado final, en el que se eliminan aquellos que no representan a un mínimo número de conceptos determinado. Este número mínimo exigido para ser un BLC válido, es el umbral mínimo del algoritmo (Izquierdo *et al.*,

2010). Mediante la combinación de diferentes valores para el umbral¸ y el tipo de relaciones consideradas (todas o solamente de hipo / hiperonimia) se obtienen diferentes conjuntos de BLC. Los *synsets* que después del proceso de filtrado quedan sin BLC asignados (debido a que su BLC sea eliminado), son procesados de nuevo para asignarles otro BLC válido en su cadena de hiperonimia. Como resultado final, se obtienen un conjunto de nuevas etiquetas capaces de conceptualizar un conjunto de sentidos.

2.4.2.1.5. SENTIWORDNET

SentiWordNet (SWN) (Esuli and Sebastiani, 2006, Baccianella et al., 2010) constituye un recurso léxico, donde se asocia cada *synset* de WN a tres puntuaciones numéricas $Obj(s), Pos(s)$ y $Neg(s)$. Cada puntuación describe cómo Objetivo, Positivo y Negativo los términos contenidos en el *synset* respectivamente. Cada una de las tres puntuaciones oscila entre {0 ... 1}, y la suma de los tres obtiene valor uno por cada *synset*. Lo que significa que un *synset* podría tener resultados distintos de cero para las tres categorías. Entonces un *synset* caracterizado por SWN tendría tres propiedades de opinión con un cierto grado (ej. *atrocious#3* $\{Pos: 0 \mid Neg: 0.625 \mid Obj: 0{,}375\}$).

La versión utilizada en este trabajo es SentiWordNet 3.0 donde se aplican dos momentos para la creación del recurso, primero un paso de aprendizaje semi-supervisado y luego un paso de recorrido aleatorio. El primer paso consiste en cuatro sub-pasos, los cuales coinciden con los aplicados en la primera versión de SWN.

- Paso de aprendizaje **semi-supervisado**
  - En el primer sub-paso, se parte de dos conjuntos semillas (uno integrado por todos los *synsets*, y contiene siete términos (paradigmáticamente positivos), y el otro integrado por todos los *synsets* que contienen siete términos (paradigmáticamente negativo) (Turney y Littman, 2003)), que se expanden a través de una serie de relaciones binarias de WordNet. Esto se hace para conectar *synsets* de una polaridad dada con otros *synsets* de la misma polaridad. La expansión de conexiones se puede realizar con un cierto radio *k*.
  - En el segundo sub-paso los dos conjuntos de *synsets* generados anteriormente se utilizan junto con otro grupo de *synsets*, donde se supone que estos tienen la propiedad Obj, y formarán parte de un conjunto de entrenamiento para la formación de un clasificador ternario (es decir, que necesita para clasificar a un *synset* como Pos, Neg o Obj). Las glosas de los *synsets* son utilizadas por el módulo de formación en lugar de los propios *synsets*, lo que significa que el clasificador resultante es de hecho un clasificador de glosa (más que un *synset*). En SWN 1.0 se utiliza un modelo de bolsa de palabras, donde la información de la glosa está representada por un conjunto de palabras que aparecen en ella (con esa vía obtienen más peso las palabras más frecuentes). En SWN 3.0 en lugar de apoyarse en las ambigüedades de glosas de forma manual, se basa en un modelo de clasificación de bolsas de *synset.* En este sub-paso todos los *synsets* (incluyendo las semillas) son clasificados por el segundo paso. En este sub-paso se mejoran los resultados al aplicar la clasificación del Paso 2 variando el radio de desplazamiento.
- Paso 2 (paso del **recorrido aleatorio**), radica en considerar a WordNet 3.0 como un grafo y se ejecuta sobre este un proceso iterativo. En este recorrido los valores de $Pos(s)$ y $Neg\ (s)$ (y, en consecuencia, $Obj\ (s)$) son nuevamente determinados en consecuencia del paso anterior. El proceso termina cuando el proceso iterativo ha convergido.

2.4.2.1.6. *EXTENDED* WORDNET

Extended WordNet (Sanda M. Harabagiu, 1999) recurso léxico creado en la Universidad de Texas. Esta propuesta se plantea mejorar la información semántica proporcionada por WN en sus distintas versiones, al agregar información semántica a las glosas y establecer nuevas relaciones entre las palabras (ahora etiquetadas a nivel de sentido) de las glosas y los *synsets* que la definen semánticamente. Esta nueva información se extrae únicamente de la parte de la definición de las glosas, al descartar los ejemplos y las aclaraciones entre paréntesis que puedan aparecer. Para la creación del recurso se aplicaron tres procesos.

- **Análisis sintáctico**. Se obtuvo mediante un proceso de *voting* aplicado con dos analizadores sintácticos. Estos se surtieron de las salidas que el etiquetador mejorado de (Brill, 1995) les propició. Esta dilatación del trayecto de análisis sintáctico no era imprescindible, pero los autores consideraron necesario un pre-procesamiento con vistas a reducir errores. El 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 *automatically* su glosa quedaría como sigue: *automatically* ***is*** *in a reflex manner***.** De este modo se realiza una asignación semántica directa entre la palabra y su glosa.
  - **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 *pure* su glosa quedaría como sigue: *pure is something not mixed.*
  - **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 *shed* su glosa quedaría como sigue: *to shed is to cast off hair, skin, horn, or feathers*.
  - **Sustantivos**. Las glosas pertenecientes a los sustantivos se extienden añadiendo el sustantivo + *is* al principio de la glosa y un punto al final de la definición. Por ejemplo, para el nombre *play* su glosa quedaría como sigue: *play is the act using a sword* (*or other weapon*) *vigorously and skillfully*.
- **Análisis lógico**. Este análisis se aplica a continuación del análisis sintáctico, se procede con una transformación donde se codifican las relaciones sintácticas del tipo sujetos sintácticos, objetos sintácticos, enlaces preposicionales, nominales complejos, y adjuntos adjetivales y adverbiales. Al utilizar las glosas conceptuales originales de WordNet, estas se transforman en su forma lógica correspondiente.
- **Análisis semántico**. En la anotación de las glosas se aplicaron dos variantes: automática y manual. La anotación automática fue aplicada bajo la supervisión de dos sistemas, uno diseñado de forma específica para desambiguar las glosas de WordNet (llamado XWN WSD) y un sistema propio para desambiguar texto libre. El proceso de decisión estuvo dado por un proceso de votación, al obtener una coincidencia entre los dos sistemas de precisión del 90 %. Se establecieron según su fiabilidad tres categorías de anotación semántica, importante resaltar que los verbos *to be* y *to have* se han tratado de forma especial y manual:
  - *GOLD*, la anotación se ha comprobado de forma manual.
  - *SILVER*, el etiquetado donde coinciden los dos métodos de WSD.
  - *NORMAL*, se ha seleccionado según la propuesta de XWN WSD.

  La creación del recurso se ha sujeto a estos procesos al seguir varias heurísticas y darle tratamiento diferenciado a variadas situaciones. Con ello se es capaz de obtener 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.

2.4.3. RECURSOS DE INTEGRACIÓN SEMÁNTICA

Con el fin de obtener información adicional para resolver diferentes problemas del PLN, se han utilizado variedad de recursos semánticos. Sin embargo, uno de los principales problemas es su descentralización. Sin embargo WN proporciona la facilidad de figurar como núcleo en el desarrollo de diferentes recursos y aplicaciones. En la actualidad algunas herramientas se han dado a la tarea de crear integraciones semánticas respetando esta idea. Se pudiera mencionar algunos trabajos como:

- **MultiWordNet**[49] (MWN) (Pianta *et al.*, 2002) es un proyecto del ITC-IRST de Trento-Italy con el objetivo de producir un WordNet italiano estrictamente alineado con WordNet de Princeton (en idioma inglés). En su primera versión, contiene alrededor de 37000 palabras en italiano organizadas en unos 28000 *synsets*, junto con información sobre la correspondencia entre el italiano y *synsets* del inglés. MultiWordNet adopta un marco metodológico distinto de EuroWordNet. Existen al menos dos modelos para la construcción de un WN multilingüe. El primer modelo, adoptado en el proyecto EuroWordNet, consiste en la construcción WN's de lenguajes específicos de forma independiente el uno del otro, al tratar en una segunda fase encontrar las correspondencias entre ellos (Vossen, 1998). El segundo modelo, adoptado en MultiWordNet, consiste en la construcción lenguajes específicos de WN's conservando en mayor medida las relaciones semánticas disponibles en el WN del inglés. Esto se hace mediante la construcción de los nuevos *synsets* en correspondencia con los *synsets* propios, y poder mantener siempre que sea posible la importación de las relaciones semánticas de los *synsets* correspondientes del WN del inglés. En MWN la información de dominios ha sido automáticamente transferida del inglés al italiano provocando la obtención de un WND para italiano además (Bentivogli *et al.*, 2004).
- **EuroWordNet** (EWN) (Dorr and Castellón, 1997, Vossen, 1998) fue desarrollado inicialmente para alinear las lenguas del inglés, español, holandés, italiano, alemán, francés, checo y estonio; y luego se realizaron nuevas versiones donde se incluye el sueco, noruego, danés, griego, portugués, vasco, catalán, rumano, lituano, ruso, búlgaro y esloveno. Para la vinculación entre los diccionarios léxicos que aquí intervienen, se utiliza el llamado ILI (Indexado Inter-Lingüístico, por sus siglas en inglés *Inter-Lingual-Index*) (Vossen *et al.*, 1999). Mediante el uso del ILI se realiza el alineamiento al tener en cuenta significados más cercanos, es decir no se logra una traslación absoluta de un lenguaje a otro para todos los sentidos incluidos en el diccionario. Esto lleva a una diferenciación de la situación de los conceptos de ILI, una reducción de la polisemia WN y una mayor conectividad entre los diferentes idiomas de WN.
- ***Meaning: Multilingual Central Repository*** (Proyecto *meaning* MCR) (Atserias *et al.*, 2004) se integra en el *framework* de EWN con cinco WNs locales incluyendo WordNet de Princeton, además aplica una versión mejorada de la ontología Concepto Superior de EWN, los Dominios de MWN, la Ontología Sugerida Superior (SUMO (en inglés *Suggested Upper Merged Ontology*)) (Zouaq *et al.*, 2009) y cientos de miles de nuevas relaciones semánticas, con las propiedades adquiridas de forma automática a partir de corpus. La primera versión de la MCR incluye solamente el conocimiento conceptual. Esto significa que únicamente las relaciones semánticas entre *synsets* han sido adquiridas de WN's locales. La actual versión de MCR integra:
    - El ILI basado en WN1.6, incluye conceptos base de EWN, la ontología de concepto superior de EWN, dominios de MultiWordNet y SUMO;

[49] http://multiwordnet.itc.it

- WN's locales (vasco, catalán, italiano y español) conectados al ILI, incluyendo WN del inglés versiones 1.5, 1.6, 1.7 y 1.7.1.
- Las grandes colecciones de preferencias semánticas, adquirida tanto en SemCor y del BNC, además de instancias incluyendo entidades nombradas (sustantivos).

2.5. MÉTODOS DE CLASIFICACIÓN

En la automatización del proceso de desambiguación se encuentran algunas dificultades, por causa de que los diccionarios presentan un alto grado de granularidad (las definiciones de las palabras son en muchos casos ligeramente diferentes), incluso lexicógrafos expertos, experimentan dificultades en las decisiones de asignación de sentidos. Para enfrentar este problema, la comunidad científica ha propuesto diversos enfoques, aunque dista de solucionarse definitivamente. Los métodos desarrollados se clasifican como métodos supervisados, débilmente supervisados, sin supervisión, basados en ejemplos de corpus, basados en conocimiento, mixtos, etc. Aunque en la actualidad solamente dos definen el camino a seguir (supervisados y no supervisados).

2.5.1. WSD CON SUPERVISIÓN

Los sistemas de WSD que se basan en métodos con supervisión (en inglés *supervised*), utilizan técnicas de máquinas de aprendizaje (como fuente de datos en un sistema de *bootstrapping*) para aprender de un clasificador de conjuntos de etiquetas[50] (clase o sentido) de entrenamiento. Estos sistemas deben disponer de grandes cantidades de datos de adiestramiento, por lo que anotar los conjuntos de entrenamientos suele ser una tarea muy laboriosa (la anotación de corpus se hace manual). En general, los enfoques de supervisión para WSD han obtenido mejores resultados que los de sin supervisión (véase la sección 2.6.1). Las técnicas supervisadas pueden ser las siguientes:

- **Listas de Decisión**, es un conjunto de reglas ordenadas (Ronald, 1987) para categorizar instancias de prueba, son vistas comúnmente como reglas "si-entonces-sino", para cada palabra *w* se lista un vector de características al que se le aplica el chequeo de la lista de decisión, este proceso acumula un valor calculado para cada sentido (David, 1994).
- **Árboles de Decisión,** es un modelo predictivo utilizado para representar reglas de clasificación con estructura de árbol, donde esta técnica particiona recursivamente el conjunto de entrenamiento. Un popular algoritmo para el aprendizaje del árbol de decisión es el C4.5 (Quinlan, 1986), una extensión del ID3 (Salzberg, 1994).
- **Clasificador Naive Bayes**, es un clasificador probabilístico basado en la aplicación del teorema de Bayes, se basa en el cálculo de la probabilidad condicional de cada sentido de una palabra a partir de un conjunto de características en el contexto (para profundizar véase (Navigli, 2009)). Autores como (Gale et al., 1992b, Mooney, 1996, Leacock et al., 1993, Escudero et al., 2000b) han hallado interesantes aplicabilidades de esta técnica en WSD.
- **Redes Neuronales**, (McCulloch and Pitts, 1988) es un grupo de neuronas artificiales interconectadas en un modelo computacional para el procesamiento de datos basado en un enfoque conexionista. Donde los pesos de enlace se adaptan progresivamente a fin de que la unidad de salida que representa la respuesta deseada, se manifieste con una

[50] Conjuntos de ejemplos codificados en términos de una serie de características, junto con su etiqueta sentido apropiado.

activación mayor que cualquier otra unidad de salida. Las redes neuronales son entrenadas hasta que la salida de la unidad que corresponda a la respuesta deseada, sea superior a la producción de cualquier otra unidad para cada formación.

- **Basado en aprendizaje de ejemplos o de instancias**, algoritmo en el que el modelo de clasificación se construye a partir de ejemplos. El modelo conserva ejemplos en la memoria como puntos en el espacio de características y ejemplos nuevos se someten a la clasificación, los que progresivamente añaden al modelo. Una de las técnicas más usadas y de mejores resultados en esta área es KNN (k-vecinos cercanos) (Navigli, 2009).
- **Máquinas de Soporte Vectorial** (*Support Vector Machines* (SVM)), introducido por (Bernhard *et al.*, 1992) se centra en la idea de aprender en un hiper-plano lineal del conjunto de entrenamiento al separar los ejemplos positivos de los ejemplos negativos.
- **Métodos de ensamblado**, cuando se desea combinar varios clasificadores se utilizan los métodos ensamblados. Se busca la combinación de las estrategias con algoritmos de diferente naturaleza. Pueden ser por **Mayoría de Votos, Mezcla Probabilística, Combinación basada en *Ranking* o Impulso Adaptativo** (*AdaBoost*).(Freund and Schapire, 1999, Escudero et al., 2000a) (para un análisis más detallado véase (Navigli, 2009)).

2.5.2. WSD DÉBILMENTE SUPERVISADO

Los sistemas de WSD que se basan en métodos débilmente supervisados (semi-supervisados, en inglés *wake supervised*), utilizan una cantidad mínima de información inicial. La diferencia entre los supervisados y estos, radica en la cantidad de información de la que parten. Los supervisados se inician desde un conjunto de información de tamaño considerable, que se utilizan para entrenar y desarrollar el sistema final. En cambio los semi-supervisados, suelen utilizar una pequeña cantidad de información inicial, para obtener un primer sistema, con el cual anotar nueva información y refinar sucesivamente dicho sistema (para obtener mayor información véase (Agirre and Edmonds, 2006)).

2.5.3. WSD SIN SUPERVISIÓN

Los sistemas de WSD que se basan en métodos sin supervisión (en inglés *unsupervised*), como principal diferencia con los supervisados y débilmente supervisados, se tiene que estos no aplican entrenamiento alguno, se basan en corpus no etiquetados y no explotan corpus manualmente anotados a nivel de sentido, evitando casi completamente la información externa. Se caracterizan por tener el potencial de superar el cuello de botella de adquisición de conocimientos (Gale *et al.*, 1992b). Este tipo de métodos se fundamentan sobre la idea, que similares sentidos de palabras tienen palabras similares asociadas. Por esta razón, se es capaz de inducir el sentido de una palabra en un texto por técnicas de agrupamiento de palabras mediante coocurrencias y luego clasificar nuevas ocurrencias en los grupos inducidos. Como característica relevante se tiene que con el uso de este tipo de métodos, no se aplica entrenamiento con el uso de texto etiquetado. En sus versiones más puras, ni siquiera hacen uso de diccionarios electrónicos, eso hace que sufran una gran desventaja, que es no poder compartir las mismas referencias (inventario de sentidos) y como consecuencia muy difícil establecer comparaciones entre sistemas.

En la tarea de desambiguación es aplicable también la discriminación de sentidos (en inglés *Word Sense Discrimination*), la cual consiste en dividir las apariciones de una palabra en un número de clases y poder determinar que pertenezcan o no al mismo sentido. Entonces, este tipo de métodos no descubre clases (sentidos) equivalentes como los sentidos que tradicionalmente se encuentran en un inventario. Esto hace también se haga muy difícil su evaluación y comparación con otras propuestas que trabajen la resolución de ambigüedad.

Las principales aproximaciones de la desambiguación sin supervisión son las siguientes:

- **Métodos basados en agrupación (*clustering*) de contexto,** en estas aproximaciones cada ocurrencia de las palabras en el corpus se representan en un vector de contexto. Los vectores son luego agrupados en conjuntos, cada uno para identificar un sentido de la palabra objetivo. Entonces, si se colocan juntos un conjunto de vectores se crearía una matriz de coocurrencia. Al tener una matriz es posible aplicarle técnicas de agrupamiento (*clustering*). Entre las más usadas está el Análisis de Semántica Latente (*Latent Semantic Análisis*[51](LSA)) (Landauer *et al.*, 1998) basada en la aplicación de la Descomposición de Valores Singulares (*Singular Value Decomposition* (*SVD*)) (Klema and Laub, 1980). Este tipo de aproximaciones se han aplicado con varias variantes de construcción de vectores de contexto, incluso utilizando la información de las glosas del los inventarios de sentidos (Navigli, 2009).
  Otro tipo de aproximaciones son las que aplican **Agrupamiento de Palabras** (*word clustering*). Estos son métodos que se orientan por agrupar palabras que son semánticamente similares, transmitiendo un significado específico. La similitud entre palabras es determinada a la información de sus características individuales, con respecto a las dependencias sintácticas que se producen en un corpus (ej. entre sujeto-verbo, verbo-objeto, adjetivo-sustantivo, etc.) (Lin, 1998a).
  Por último las aproximaciones por **Grafos de Coocurrencia,** estos enfoques se basan en la noción de un gráfico de co-ocurrencia, es decir, un grafo $G = (V, E)$ cuyos vértices corresponden a las palabras en un texto y conexiones *E* entre pares de palabras que co-ocurren en una relación sintáctica, en el mismo párrafo, o en un contexto más amplio. La construcción de un grafo sobre la base de co-ocurrencia relaciones gramaticales entre las palabras en su contexto se describió por (Widdows and Dorow, 2002).
- Otro tipo de distinción son los **métodos basados en conocimiento** (rico en conocimiento o basado en diccionario) y los basados en ejemplos de corpus (o pobre en conocimiento). Los primeros se basan en el uso de recursos externos de léxico, como comprensibles por máquina, como son los diccionarios, tesauros, ontologías, etc. Mientras que el segundo no hace uso de estos recursos para la desambiguación, solamente utiliza informaciones extraídas de los corpus.

Luego de haber observado estas caracterizaciones generales se puede decir que el problema de WSD se puede tratar además de dos maneras distintas, Basada en *Tokens* (*token-based*) y Basada en Tipo (*type-based*). La primera es capaz de asociar un significado específico para cada ocurrencia de la palabra en correspondencia del contexto donde esta aparezca. Y *type-based* está enfocado por el sentido predominante (*predominant sense*) sobre el análisis del texto en su totalidad.

Finalmente se encuentran los métodos mixtos que combinan varios métodos con el fin de la desambiguación léxica. Hasta este punto del epígrafe se ha mostrado una caracterización general de las aproximaciones de WSD, es importante destacar que dentro de cada una de ellas existen disímiles variantes.

Debido a los grandes requerimientos de recursos anotados manualmente que se necesitan para aplicar la desambiguación mediante las propuestas supervisadas. En esta Tesis se le da un voto de confianza a la aplicabilidad de métodos sin supervisión, basados en conocimiento, con el objetivo de propiciar un Análisis Semántico Multidimensional (donde intervienen varias bases de conocimiento) en las tareas de PLN en especial WSD.

[51] http://onlinelibrary.wiley.com/doi/10.1002/aris.1440380105/pdf

2.5.4. MÉTODOS BASADOS EN CONOCIMIENTO

Este tipo de métodos necesitan de bases de conocimientos para realizar sus análisis y emitir sus respuestas. Dígase tesauros, lexicones, diccionarios u ontologías. En esta categoría se encuentran diferentes algoritmos para la etiquetación automática de sentidos. Normalmente, el rendimiento 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, pueden aplicarse a cualquier tipo de texto en comparación con los basados en corpus, que solamente se pueden aplicar a aquellas palabras de las que disponen corpus anotados (Vázquez, 2009).

A la par del surgimiento de las bases de conocimiento léxicas, se han desarrollado diferentes técnicas utilizadas por los métodos basados en conocimiento, entre las bases de conocimiento más utilizadas en WSD está WordNet (Miller et al., 1990). A continuación se enumeran algunos tipos diferentes de métodos basados en conocimiento que más adelante se describen:

El **algoritmo de Lesk** (Lesk, 1986), en el cual las definiciones de los sentidos de las palabras en un contexto se solapan obteniendo una medida de similitud.

- Aproximaciones estructurales
    - **Medidas de similitud semántica** extraídas a través de redes semánticas. Estas medidas incluyen métodos que definen distancias existentes entre conceptos de una red semántica. Debido a que los procesos de desambiguación dependen del contexto, si este se hace muy grande varían las características extraídas de él, por ello es que en dependencia del tamaño del contexto estas medidas se dividen en dos grandes categorías:
        - Aplicado a contextos locales
        - Métodos aplicables a contextos globales
    - Aproximaciones basadas en Grafos
- **Preferencias de selección** adquiridas de forma automática o semiautomática, como una forma de restringir los posibles sentidos de una palabra, basados en la relación que esta tiene con otras palabras en el contexto.
- **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

A continuación se describirán algunos métodos basados en conocimiento en correspondencia con los distintos grupos en los que se dividen.

2.5.4.1. ALGORITMO DE LESK Y SUS VARIACIONES

En 1986 Lesk propone un algoritmo basado en el conocimiento contenido en el *Diccionario Oxford Avanzado*[52]. Para explicarlo se tomará el ejemplo ilustrado del artículo (Lesk, 1986), donde se tienen las definiciones de la palabra *pine* y la palabra *cone*, además de presentar como contexto que estas se encuentren juntas. Por ejemplo:

pine

1. *"* seven kinds of evergreen tree with needle-shaped leaves."*
2. *" pine."*

[52] http://www.oed.com

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*)*."*

En el ejemplo anterior se determinan que los sentidos que más se solapan son el primero de *pine* y el tercero de *cone*. Esta propuesta aunque parece sencilla ha servido sirvió como base para muchas otras que se denominan variaciones de Lesk. Aunque se considera basada en diccionario constituye un punto de partida para aproximaciones basadas en corpus, ya que el objetivo se centra en el solapamiento contextual. Este algoritmo ha sido evaluado sobre el corpus de Senseval-1 obteniendo un valor de precisión y cobertura alrededor de 55% (Kilgarriff and Rosenzweig, 2000) en la tarea *Lexical Sample*. También ha sido probado sobre Senseval-2 en la tarea *English All Words* con la obtención de un 42% de precisión y cobertura aproximadamente (Vasilescu *et al.*, 2004).

Una de las dificultades de Lesk, es que depende en gran medida de la exactitud de las palabras que se encuentran en las definiciones, ya que el solapamiento se establece al buscar coincidencias exactas entre ellas. Otros de los problemas es la explosión combinatoria, pues por lo general, cada palabra en un contexto o frase tiene varias definiciones. Por ejemplo:

"The use (*13*) *of paper* (*10*) *and pencil* (*5*) *seems* (*4*) *to be* (*14*) *the most* (*5*) *natural* (*13*) *way* (*13*) *to create* (*6*) *concept* (*1*) *maps* (*8*)*."*

Como se observa al lado de cada palabra se enumera el total de definiciones de cada una, entonces para determinar el sentido correcto de todas, se requiere de 1476384000 iteraciones. Esto sucede porque se combina cada definición de la palabra, con cada una de las definiciones de las que se consideran vecinas.

Para dar solución a este problema (Cowie *et al.*, 1992) propone ***Simulated Annealing***, donde se define una función que refleja las combinaciones de sentidos en un texto, y cuyo valor mínimo se corresponde con la selección de los sentidos correctos. El objetivo, es encontrar la combinación de sentidos que minimiza esta función. Para este propósito, se determina una combinación inicial de sentidos (por ejemplo, se recogen los más frecuentes para cada palabra), y entonces se realizan varias iteraciones, donde la definición de una palabra aleatoria en el texto se reemplaza con otra distinta, y la nueva selección se considera correcta únicamente si reduce el valor de la función. Las iteraciones terminan cuando no existe ningún cambio en la configuración de los sentidos (Vázquez, 2009). Este método consigue un 47% de precisión a nivel de sentidos y luego en el 2001 fue reimplementado obteniendo un valor superior de precisión (65.24%) en un corpus etiquetado con los sentidos del diccionario LDOCE.

Otro intento de mejorar el algoritmo de Lesk lo propone **Wilks y Stevenson** (Wilks and Stevenson, 1996) donde introduce la obtención de un **vector de coocurrencia** (2) entre las definiciones de las palabras y el contexto en que se localizan, donde logra mostrar la frecuencia de aparición en cada una de las definiciones (ecuación (3)). Los vectores resultantes serían posteriormente evaluados para determinar sus cercanías o similitudes con el vector del contexto y para ello se podría utilizar la ecuación matemática del coseno del ángulo (ecuación (4)) entre vectores. Las pruebas realizadas de esta nueva propuesta se hicieron también sobre LDOCE, en este diccionario existen un número de palabras reducido (2181 palabras) las cuales acotan las frecuencias de las mismas. Para evaluar la propuesta se tomó la palabra *bank* alcanzando se un 45% de exactitud en la identificación del sentido y un 90% en la detección de palabras homófonas.

$$\vec{v}_w = (v_0^w, v_1^w, v_2^w, \dots, v_N^w) \quad (2)$$

$$v_i^w = f_{w,Z_i} \quad (3)$$

$$cos(\vec{v},\vec{u}) = (\frac{\sum_{i=1}(v_i * u_i)}{\sqrt{\sum_{i=1} v_i^2} * \sqrt{\sum_{i=1} u_i^2}}) \quad (4)$$

Una nueva versión del algoritmo de Lesk, que también trata de resolver el problema de la explosión combinatoria, es el **Algoritmo de Lesk Simplificado** (Kilgarriff and Rosenzweig, 2000) que utiliza un proceso separado de desambiguación para cada palabra del texto de entrada. En esta propuesta el sentido correcto se determina individualmente. La aproximación toma cada palabra de forma individual, sin tener en cuenta el sentido de las otras que aparecen junto a ella. Esta variante agiliza el proceso y aporta mejores resultados que el original. Los pasos a seguir son los siguientes:

- Determinar el solapamiento para cada sentido *s* de cada palabra *w*, que consiste en el número de palabras en común entre la definición del sentido *s* y el contexto donde aparece la palabra.
- Encontrar el sentido *s* con el máximo solapamiento
- Asignar el sentido *s* a *w*

Es de destacar que en esta propuesta el solapamiento no se realiza entre definiciones de palabras, sino que se busca medir similitud entre definiciones y frase contextual donde aparece la palabra. Esta propuesta ha sido ejecutada por (Vasilescu *et al.*, 2004) sobre el corpus de Senseval-2 con resultados del 57.66% de exactitud en la cobertura y un 58.18% de precisión, los que demuestran que la variante simplificada supera a la original.

Otra versión de Lesk se conoce como **Algoritmo de Lesk basado en Corpus** (Kilgarriff and Rosenzweig, 2000), aumenta el contexto de una palabra con ejemplos adicionales etiquetados en corpus anotados para resolver la ambigüedad. En ella, se mantiene seleccionado aquel sentido que tenga mayor solapamiento con algún contexto pre-etiquetado. Pero la vía que elije para la asignación de pesos es asignar la inversa de la frecuencia. Esta aproximación ha logrado mejores resultados que los anteriores. Comúnmente esta propuesta es comparada con las supervisadas debido a su extensión de búsqueda de ejemplos anotados en corpus. Aunque Lesk basado en corpus no explícitamente representan las frecuencias relativas de las etiquetas de sentido de corpus. Sí implícitamente favorece etiquetas comunes, debido a que estos tienen grandes conjuntos de contexto, y una palabra arbitraria en una frase del corpus de prueba es más probable que ocurra en el contexto de una etiqueta de conjunto. Este algoritmo según las corridas aplicadas por (Kilgarriff and Rosenzweig, 2000) sobre Senseval-1 obtuvieron resultados de precisión y cobertura alrededor del 69% en la tarea *Lexical Sample*.

También se ha implementado lo que se conoce como Algoritmo de Lesk Adaptado, desarrollado por (Banerjee and Pedersen, 2002). Ofrecen una nueva propuesta de **Espacios Semánticos Aumentados**, donde introducen el uso de WN (Miller et al., 1990), ya sea la utilización de sus conceptos como sus relaciones, es decir, tal y como se hacía en *Simulated Annealing* (Cowie *et al.*, 1992) al buscar la definición de mayor solapamiento. Ahora se analiza el sentido con mayores solapamientos de conceptos asociados mediante relaciones semánticas. Un ejemplo tomado de (Banerjee and Pedersen, 2002) se ilustra en la Figura 6 donde se expande un *synset*:

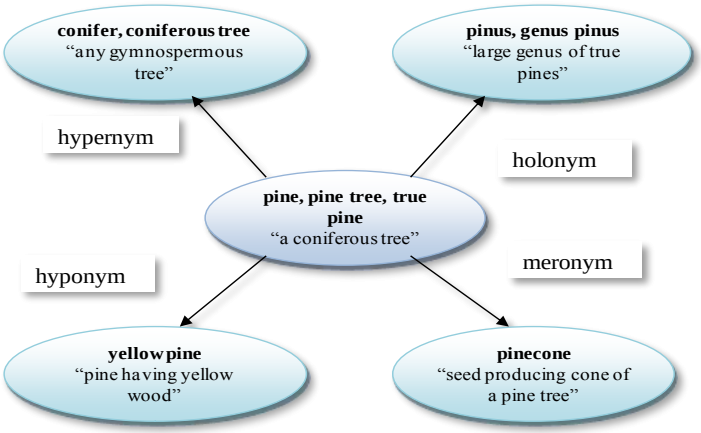


Figura 6: Conceptos de WordNet asociados al sentido *pine#1*.

Algo parecido a esta propuesta se encuentra la planteada por (Richardson, 1997) el cual presentó como contexto una red semántica extraída de LDOCE y *Webter's 7th Gove*, mediante el análisis sintáctico de sus definiciones. El objetivo se centra en extraer relaciones semánticas y asignarles peso según la frecuencia. Debido a que estos diccionarios no están etiquetados a nivel de sentidos, lo que se logra es establecer relaciones entre palabras (según la definición toma sentido).

2.5.4.2. APROXIMACIONES ESTRUCTURALES

Debido a la disponibilidad de recursos léxicos computacionales como WordNet y otros ya descritos, se han generado una serie de enfoques estructurales con el fin de sacar provecho de las estructuras que en las redes semánticas se encuentran. Entre las estrategias que se han seguido, está el reconocimiento y valoración de los patrones semánticos y estructurales que se pueden obtener con el uso de este tipo de recursos tanto en un contexto local o global. Se presentan dos enfoques principales, los métodos basados en las similitudes semánticas y basados en grafos. Aunque se pueden encontrar propuestas donde se interrelacionen ambos enfoques, por ejemplo en (Sinha and Mihalcea, 2007).

2.5.4.2.1. MEDIDAS DE SIMILITUD SEMÁNTICA

La similitud entre palabras o conceptos busca cuantificar el grado de cercanía al utilizar relaciones semánticas, en su mayoría aplican a redes semánticas (ej. WN). A continuación se muestran una serie de medidas de similitud ya experimentadas sobre WN, que contienen aspectos en común con la propuesta de esta investigación. Por lo general, estas toman un par de palabras de entrada y devuelven un valor que indica el grado de similitud que existe entre ambas palabras.

- (Leacock and Chodorow, 1998) proponen la ecuación (5), sustentada en el camino mínimo entre las dos palabras de entrada, devolviendo un valor normalizado según la profundidad de la taxonomía[53].

[53] Un listado de tópicos o categorías, usualmente jerárquico (Relaciones padre-hijo). No necesariamente incluye definiciones. Puede incorporar contenido tanto de tesauros como de ontologías.

$$Similitud(C1, C2) = -\log(\frac{Camino(C1, C2)}{2D}) \quad (5)$$

Donde $Camino\ (C1, C2)$ representa el número de arcos que conecta a los dos conceptos y D es la profundidad total de la taxonomía.

También proponen la ecuación (6) donde C y k son constantes, el $Camino\ (C1, C2)$ se define de la misma forma que en la ecuación (5) y d representa el número de cambios de dirección.

$$Similitud\ (C1, C2) = C - Camino\ (C1, C2) - kd \quad (6)$$

- En (Resnik, 1995) se introduce 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.

$$IC(C) = -log(P(C)) \quad (7)$$

Dado un corpus, $P(C)$ es la probabilidad de encontrar una instancia de tipo C. El valor para $P(C)$ es mayor en 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 una única cima, entonces el valor para este concepto es uno). Resnik define una medida de similitud semántica entre dos palabras al utilizar 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 que existe un camino desde la palabra $w1$ y la palabra $w2$. En la ecuación (8) se muestra esta medida.

$$Similitud(C1, C2) = IC\ (LCS(C1, C2)) \quad (8)$$

- En (Jiang and Conrath, 1997) presentan una alternativa a la medida de Resnik al utilizar la diferencia existente en el contenido de información de los dos conceptos para indicar su similitud. (Véase la ecuación (9)).

$$Similitud(C1, C2) = 2 * IC(LCS(C1, C2)) - (IC(C1) + IC(C2)) \quad (9)$$

- (Mihalcea and Moldovan, 1999) introducen una nueva ecuación que sirve para medir la similitud entre jerarquías independientes. Con esta medida, crean caminos virtuales entre ellas a través de las definiciones de las glosas en WordNet. En la ecuación (10) $descendientes\ (C2)$ es el número de conceptos en la jerarquía de $C2$, y Wk es un peso asociado con cada concepto (este valor se identifica como su profundidad dentro de la jerarquía), $|CD_{12}|$ es el número de palabras comunes en las definiciones.

$$Similitud\ (C1, C2) = \frac{\sum_{k=1}^{|CD_{12}|} Wk}{\log\ (descendientes(C2))} \quad (10)$$

- ***Conceptual Density*** **(Densidad Conceptual)** es un término propuesto por (Agirre and Rigau, 1996), donde se busca solapar una jerarquía C con las palabras del contexto de C. Esta se corresponde con la ecuación (11).

$$CD(C, m) = \frac{\sum_{i=1}^{m} W_i}{\log\ (descendientes\ \ (C))}, \text{ Donde } W_i = nhyp^{i^{0.20}} \quad (11)$$

Donde m es el total de sentidos a desambiguar, $descendientes(C)$ corresponde al número de conceptos en la jerarquía enraizada por C. Wk es un peso para cada concepto

en la jerarquía calculado por $nhyp$, el cual representa es el número de hipónimos de un nodo i (suavizado por el valor determinado empíricamente a 0.20). El método de WSD que utiliza esta medida elige el sentido de mayor solapamiento, es decir el de mayor densidad conceptual (Véase la Figura 7).

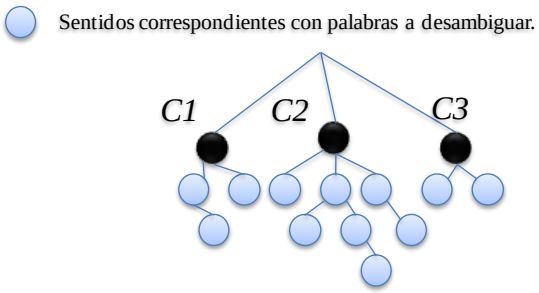


Figura 7. Ejemplo de densidad conceptual.

- El cálculo de **Información Mutua** (*Mutual Information* (MI) (Church and Hanks, 1990) se basa en la coocurrencia de las palabras y trabaja con la frecuencia de aparición de dos palabras juntas.

$$MI(x,y) = log_2(\frac{P(x,y) * N}{P(x) * P(y)}) \quad (12)$$

Donde $P\ (x,y)$ indica la probabilidad de coocurrencia de que x esté junto a y, y la $P\ (x)$ y $P\ (y)$ que se encuentren en el documento, N es el total de eventos.

Si $MI\ (x,y) > 0$: la palabra x, aparece junto a la palabra y más a menudo de lo que a simple vista podría parecer.

Si $MI\ (x;y) < 0$. Significa que la palabra x, aparece junto a la palabra y menos frecuentemente.

Si $MI(x;y) \approx 0$. Indica que no existe ninguna evidencia que las relacionen.

- A esta fórmula se le han hecho adaptaciones en su interpretación (Church and Hanks, 1990, Pekar and Krkoska, 2003), para aplicarlo a jerarquías taxonómicas. De forma que una de sus variantes se expone en la ecuación (13).

$$MI(w,D) = log_2\left(\frac{P(w,D)}{P(w)*P(D)}\right) = log_2(\frac{P(D,w)}{P(D)}) = log_2(\frac{P(w,D)}{P(w)}) \quad (13)$$

Donde *w* (palabra), *D* (dominio) y *P* es la probabilidad.

- Una nueva forma de establecer la relevancia de una palabra *w* sobre un dominio D, se mide al utilizar Radio de Asociación (*Association Ratio*) (AR) (Rigau Claramunt, 1998) (Véase la ecuación (14).

$$AR(w,D) = P(w,D) * log_2(\frac{P(w,D)}{P(w)}) \quad (14)$$

En este caso no solamente se utilizan las relaciones semánticas de WordNet, sino que su principal plataforma es utilizar las relación que proporcionan los mapeos con WND (Sara and Daniele, 2009). Al aplicar la fórmula de AR se cuantifica la relación de las palabras del diccionario de WN y los conceptos de la jerarquía de WordNet *Domains*. Este método es aplicable a cualquier mapeo de WordNet incluso dentro del propio WN (Vázquez *et al.*, 2004a).

- Otras de las medias aplicadas en recursos estructurales (ej. WordNet) son la propuestas por (Lin, 1998b). Las ecuaciones (15) y (16) son distancias de edición utilizadas para la

comparación de las definiciones de los *synsets* de WN, solamente la ecuación (17) aplica métodos probabilísticos.

$$Sim_{edit}\ (x,y) = \frac{1}{1+editDistance\ \ (x,y)} \tag{15}$$

Donde $editDistance$ obtiene el mínimo número de operaciones de inserción y eliminación para transformar una cadena en otra. Para calcular la distancia de edición se pueden aplicar variablemente cualquier medida de distancia de edición (ej. Levenshtein[54]).

La segunda distancia de Lin corresponde con la ecuación (16). Esta ecuación se basa en el análisis de diferentes trigramas.

$$Sim_{tri}(\text{x,y}) = \frac{1}{1+|tri\,(x)|+|tri\,(y)|-2*|tri\,(x)\cap tri\,(y)|} \tag{16}$$

Donde $tri(x)$ corresponde al conjunto de trigramas en x a nivel de caracteres (ej. $tri(home) = \{hom, ome\}$).

Y la tercera medida de similitud aplica trigramas de modo probabilístico.

$$Sim_{tri}(\text{x,y}) = \frac{2 \times \sum_{t\ \in tri\,(x)\cap tri\,(y)} log\ P(t)}{\sum_{t\ \in tri\,(x)} log\ P(t)+\sum_{t\ \in tri\,(y)} log\ P(t)} \tag{17}$$

- Entre las tantas medidas que se han adaptado también se puede hablar de la obtención de ***Reuters Vector*** **(**ecuación (18)), intentando cuantificar la probabilidad de acercamiento entre los sentidos de las palabras y los dominios de WND (Magnini *et al.*, 2002).

$$P(D|s)\ = \frac{P(s|D) * P(D)}{\sum_{i=1}^{n} P(s|D_i) * P(D_i)} \tag{18}$$

$$\text{Donde } P(s|D) = \ P(l_1, l_2, \dots, l_m|D) = \prod_{i=1}^{m} P(l_i|D)$$

Donde D es el dominio de WND y s es el *synset* asociado, P es la probabilidad.

$$P(l_i|D) = \frac{c(l_i, D) + \in}{c(D) + \in |L|} \tag{19}$$

Tal que $c\ (l_i, D)$ es el número de coocurrencias del lema en el dominio, $c\ (D)$ total de dominios, $|L|$ es el número de lemas del corpus y $\in$ es la constante de Épsilon.

- Una de las medidas más usadas hoy en día la proponen (Cilibrasi and Vitányi, 2007), la cual se denomina **Distancia Normalizada de Google**[55] (en inglés, *normalized Google distance*), la ecuación (20) representa dicha distancia. Donde N es el número total de páginas Web, y al igual que la ecuación (17) se busca cuantificar la cercanía entre dos palabras, define la frecuencia $f(x)$.

  Si $f(x), f(y) > 0$ & $f(x,y) = 0$ entonces $NGD(x,y) = \infty$ se analiza lo siguiente: $NGD(x,y)$ está indefinida si $f(x) = f(y) = 0$;

[54] http://www.miislita.com/searchito/levenshtein-edit-distance.html

[55] Para más detalles consultar Cilibrasi, R. L. & Vitányi, P. M. B. (2007) The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, VOL. 19, NO 3.

$NGD(x,y) = \infty$ para $f(x,y) = 0$ y cualquiera o ambos $f(x) > 0$ & $f(y) > 0$;
$NGD(x,y) = 0$ otro caso.

Este algoritmo es aplicable en las redes semánticas siempre que se haga ciertas consideraciones.

$$NGD\ (x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (20)$$

Algunas de las medidas aquí descritas han sido centralizadas en una única herramienta llamada *WordNet Similarity* [56]. Por ejemplo las propuestas por (Leacock and Chodorow, 1998, Jiang and Conrath, 1997, Resnik, 1995, Lin, 1998b) y la propuesta de (Banerjee and Pedersen, 2002) fueron incluidas en dicho recurso además de otras que mantienen la misma filosofía.

Existen algunas propuestas del PLN que no precisamente son una medida, pero la aplicación de sus algoritmos consigue obtener valores y elementos estructurales válidos para procesos semánticos. Se pudiera mencionar **Marcas de Especificidad** (ME) (Montoyo Guijarro, 2002). Este método explora en el interior de la jerarquía de WordNet para determinar conceptos comunes de un conjunto de palabras y aplicarlo en WSD. Al utilizar la taxonomía de nombres de WN y sus relaciones de hiponimia e hiperonimia, es capaz de desambiguar palabras dentro de un contexto local (oración). Similar a esta idea se pueden ver las propuestas de creación de clases semánticas como por ejemplo (Izquierdo *et al.*, 2007), donde al conocer la taxonomía es posible generar o detectar elementos ancestros que reducen el grado de granularidad de un recurso (ej. WordNet) sin perder su significado.

2.5.4.2.2. APROXIMACIONES BASADAS EN GRAFOS

En los años cercanos al 1976 surge una nueva visión del análisis textual defendido por (Halliday and Hasan, 1976). Donde introduce el término de "cohesión" entre los recursos necesarios para la construcción del texto y la gama de significados que están relacionados específicamente con relación con lo que se habla o por escrito a su entorno semántico. Para concretar, este estudio se centra en el análisis de la cohesión que se deriva de las relaciones semánticas entre las oraciones. Como son: referencias de unos a otros, la repetición de significados de palabras, etc. Se puede decir que muchos de los trabajos científicos realizados sobre esta línea, se han visto inspirados en la generación de cadenas léxicas (en inglés *lexical chains*), las que toman como base las ideas de (Halliday and Hasan, 1976). La creación de cadenas léxicas consiste en obtener un secuencia de palabras semánticamente relacionadas desde $w_1, w_2,...,w_n$. pertenecientes a un texto, donde w_i se relaciona léxico-semánticamente con w_{i+1}. Por ejemplo mediante las relaciones es-un, es-parte-de, tiene-como-parte y otras. Las cadenas léxicas determinan los contextos y contribuyen a la continuidad del significado con el ofrecimiento de la coherencia de un discurso (Véase el ejemplo de la Figura 8).

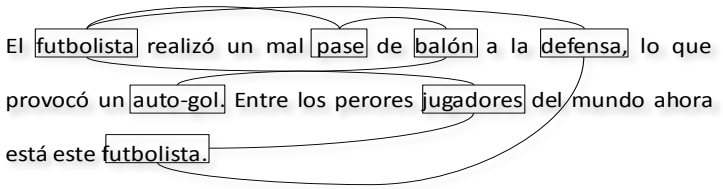


Figura 8. Cadenas léxicas obtenidas en un pequeño discurso.

[56] Disponible en http://search.cpan.org/dist/WordNet-Similarity/

Como se puede apreciar en el ejemplo de la Figura 8 se relacionan las palabras encerradas, para lograr establecer la semántica del contexto en alguna medida. La generación de este tipo de estructuras se ha aplicado en el área de cohesión de texto. Pero además de esta área, también ha sido empleada en la difícil tarea de elaboración de resúmenes automáticos (en inglés *text summarization*) (Barzilay and Elhadad, 1997, Oliveira, 2006), en la extracción de palabras claves (Ercan and Cicekli, 2007) y otras.

De modo general el algoritmo de creación de cadenas léxicas procede de la siguiente forma:

1. Se seleccionan las palabras candidatas del texto (la mayoría pertenecen a la misma categoría léxica).
2. Para cada palabra candidata, 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.
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. Un ejemplo de generación de cadenas léxicas tomado de (Vázquez, 2009) se puede ver en la Figura 9 donde se logran alinear tres palabras de un contexto mediante sus definiciones.

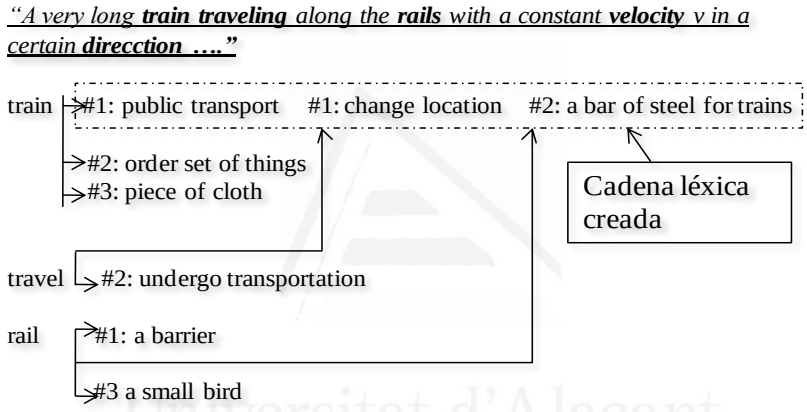


Figura 9. Ejemplo de creación de una cadena léxica a partir de sentidos.

Debido a que con el uso de cadenas léxicas se establecen relaciones semánticas a nivel estructural, esto permite el poder aplicar las medidas de similitud comentadas en la sección 2.5.4.2.1. La primera medida aplicada en esta temática fue introducida por (HIRST and ST-ONGE, 1998), donde plantea medir la fortaleza entre dos sentidos de dos palabras S_{w1} y S_{w2}, al aplicar la siguiente ecuación:

$$scoreHso(Sw_1, Sw_2) = C - d(Sw_1, Sw_2) - k * turns(Sw_1, Sw_2) \quad (21)$$

Donde C y k son constantes, d representa la distancia mínima entre los dos sentidos en la Taxonomía de WordNet y $turns$ es el número de veces que la cadena cambia de dirección. Un cambio de dirección no es más que el uso de una relación inversa (ej. tipo-de, tiene-tipo). Se propone analizar un ejemplo donde se pretende elegir de los posibles sentidos de las palabras vehículo y aeronave, los relacionados con mayor fortaleza. Por ejemplo, en la Figura 10 en auto aparece un cambio de dirección. Al aplicar la ecuación (21) entre auto o aeronave esta quedaría de la siguiente manera:

$$scoreHso(auto, aeronave_1^{n'}) = C - 4 - k * 1 \quad (22)$$

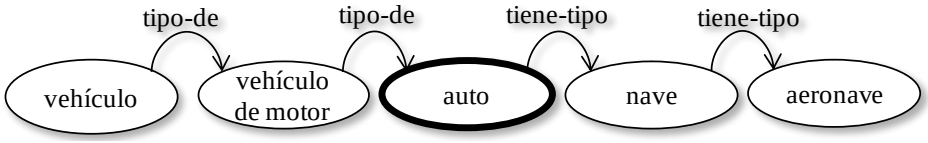


Figura 10. Ejemplo de cambio de dirección en una cadena léxica.

Como se puede observar, este tipo de algoritmos para el cálculo de las cadenas léxicas a menudo realizan desambiguación antes de inferir que las palabras se relacionan semánticamente. Esta propuesta sufre de desambiguación inexacta, debido a que desambigua palabras a la primera vez que la encuentra.

Propuestas como las de (Barzilay and Elhadad, 1997), se refieren a resolver de alguna forma la inexactitud del enfoque original, donde mantiene todas las interpretaciones posibles hasta que todas las palabras que se van encadenando hayan sido consideradas. Lo cual introduce nuevas complicaciones, debido al tratamiento de muchas combinaciones posibles de sentidos de las palabras en el texto. Luego, esta propuesta es enmarcada en un algoritmo en tiempo lineal propuesto por (Silber and McCoy, 2000) que supera los resultados del al anterior.

Basado en esos trabajos, (Galley and McKeown, 2003) desarrollaron un método de desambiguación que consta de dos pasos: Primero construyen un grafo que representa todas las posibles interpretaciones de las palabras, estableciendo conexiones entre cada palabra contra todas las palabras previamente leídas. Estas son representadas como nodos en un grafo y las relaciones semánticas como arcos. Luego, en el proceso de desambiguación, se agrupan todas las ocurrencias de la palabra obtenida y entonces para cada sentido de la palabra objetivo, se suman las fortalezas calculadas en sus conexiones. Como resultado final se elige el sentido de mayor puntuación. Esta propuesta reportó un 62,1% de exactitud en la desambiguación de los sustantivos al ser aplicado sobre un subconjunto del corpus de SemCor.

Varias aproximaciones han propuesto construir grafos con la información semántica existente en un texto. Por ejemplo, aquellas que usan las Interconexiones Semánticas Estructurales (en inglés *Structural Semantic Interconnections* (SSI)) (Navigli and Velardi, 2005) capaces de crear especificaciones estructurales para cada posible sentido de cada palabra en el contexto. En primer lugar, dado un contexto de la palabra *w*, SSI construye un sub-grafo del léxico WordNet que incluye todos los sentidos de las palabras en *w* y conceptos intermedios que se producen. Esto genera una cadena léxica válida para conectar un par de sentidos en *w*. En segundo lugar, el algoritmo selecciona los sentidos de las palabras en su contexto, al buscar maximizar el grado de conectividad del sub-grafo inducido del léxico WordNet.

Otras propuestas ofrecen la capacidad de explorar la integración de WN y FrameNet (Laparra et al., 2010, Navigli and Velardi, 2004) y otras más que se han aventurado a aplicar el uso del conocido algoritmo PageRank (implementado por el buscador de Google) como son (Agirre and Soroa, 2009, Reddy et al., 2010), (Soroa et al., 2010, Sinha and Mihalcea, 2007) basándose en las informaciones ofrecidas por las interconexiones de la Base de Conocimiento Léxica de WN (en inglés *Lexical knowledge Base* (LKB) (incluye *eXtended* WordNet)). Todas estas aproximaciones basadas en grafos, tienen la atención de la comunidad científica mundial, debido a los prometedores resultados que han publicado. Por ejemplo, (Sinha and Mihalcea, 2007) y (Agirre and Soroa, 2009) obtuvieron en experimentaciones sobre el corpus de *test* de la competición Senseval-2 56.37% y 58.6% de exactitud respectivamente para *All Words*, resultados que los colocarían en los primeros puestos del *ranking* de esa competición.

2.5.4.3. 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 (ej. comer-comida, beber-líquidos). Estas se consideran son ejemplos de restricciones semánticas y pueden ser utilizadas para desechar sentidos incorrectos y seleccionar solamente aquellos que se corresponden con las reglas. Se pueden tener en cuenta dos puntos de vista, la generación de un conjunto de reglas de restricción del número de posibles sentidos, y segundo seleccionar aquellos sentidos que mejor satisfagan un conjunto de restricciones.

El tipo de relación a tener en consideración no queda únicamente en el hecho sintáctico del ejemplo, sino que se pueden establecer relaciones del tipo palabra–clase y clase-clase. Con estas nuevas relaciones es posible aliviar el problema de escasez de datos y taxonomías elaboradas manualmente como WN. (Agirre and Martinez, 2001) establecieron comparaciones puntuales entre los tres modos de aplicar la preferencia de selección (ej. palabra-palabra, palabra-clase, y clase-clase). El cálculo de relaciones palabra-clase según (Resnik, 1993) se infiere mediante la cuantificación de la contribución de una clase semántica al utilizar todos los conceptos que comparten esa clase. Para una explicación más extensa de relación palabra-clase (véase (Resnik, 1993)). En las relaciones entre palabras se utiliza la ecuación (24), con el fin de determinar cuántas veces co-ocurre la palabra w_1 con la palabra w_2 mediante la relación R. Como consecuencia, se puede determinar la probabilidad de aparición de la relación entre dos palabras, con respecto a la probabilidad de aparición de esa misma relación R, con respecto a una de las dos palabras en todo el corpus (véase la ecuación (24)).

$$Cont\ frec(w_1, w_2, R) \tag{23}$$

$$P(w_1, w_2, R) = \frac{Cont\ frec(W_1, W_2)}{Cont\ frec(W_1, R)} \tag{24}$$

2.5.4.4. MÉTODOS HEURÍSTICOS,

Una forma fácil de resolver la ambigüedad semántica de las palabras es utilizar heurísticas basadas en propiedades lingüísticas inducidas luego del análisis de textos extensos. Una de las heurísticas más utilizadas como sistema de comparación básico (en inglés *baseline*) es da determinación del sentido más frecuente. Otras dos heurísticas comúnmente utilizadas son un sentido por discurso y un sentido por colocación.

2.5.4.4.1. EL SENTIDO MÁS FRECUENTE

Esta heurística denominada *Most Frequency Sense* (MFS) se centra en las propiedades que tiene el lenguaje que por naturaleza utilizar palabras (ciertos significados o definición) con más frecuencia que otras. Un sistema muy simple de desambiguación sería aquel que asignara a cada palabra su sentido más frecuente. Es preciso decir que el diccionario léxico de WordNet ordena sus *synsets* a partir de la frecuencia que estos presentan en el corpus de SemCor[57]. Debido a ello,

[57] http://www.cse.unt.edu/~rada/downloads.html#semcor

disímiles aproximaciones ayudan a sus sistemas desambiguación con la obtención del primer sentido de WN (Navigli, 2009). (Mc.Carthy *et al.*, 2004) demuestra cómo determinar el sentido predominante en un dominio determinado, al aplicar medidas de similitud entre distintos sentidos de una palabra y palabras similares. Esta propuesta obtuvo una exactitud de un 64% sobre el corpus del *test* de Senseval-2 de la tarea *English All Words* para los sustantivos. El algoritmo utilizado por este método se compone los siguientes momentos:

Dada una palabra *w* encontrar las $top\ k$ (k mayores) palabras similares a partir del análisis de corpus.

$Nw = \{n_1, n_2, \ldots, n_k\}$ con sus respectivos valores de similitud

$\{S(w,n_1), S(w,n_2), \ldots, S(w,n_k)\}$

Para cada sentido ws_i de w, identificar la similitud con las palabras n_j al usar el sentido de n_j que maximice el valor de similitud. El *ranking* de sentidos *ws* se establece al ordenarlos basándose en el valor de similitud total.

$$S(ws_i) = \sum_{n_j \in N_w} S(w,n_j) \frac{Sw(ws_i,n_j)}{\sum_{ws_i \in Sentidos\ (w)} Sw(ws_i,n_j)} \quad (25)$$

$$\text{Donde, } Sw(ws_i,n_j) = max_{\ n_x \in\ Sentidos\ (w_j)}\ Sw(ws_i,ns_x) \quad (26)$$

A continuación se muestra un ejemplo tomado de (Vázquez, 2009) donde se desea determinar el sentido de la palabra ***pipe*** en un texto determinado. Los posibles sentidos de ***pipe*** para el inglés son:

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

Las palabras similares detectadas en el texto son las siguientes:

$$N=\{tube, cable, wire, tank, hole, cylinder, fitting, \ldots\}$$

Para cada palabra, *N* se calcula en valor de similitud con el sentido ***pipe#i*** (al escoger el valor de similitud que maximiza el par).

- *pipe#1 - tube#3 = 0.3*
- *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*

Es importante destacar que esta aproximación obtuvo en la competición de Senseval-2 un resultado destacado, con valores de precisión de 64% en la tarea de *all nouns*.

2.5.4.4.2. UN SENTIDO POR COLOCACIÓN

Introducido por (Yarowsky, 1993), se plantea que una palabra tiende a tener el mismo sentido cuando se utiliza en la misma colocación (es decir en un lugar donde se rodea de similares palabras). Entre las colocaciones más efectivas está la adyacencia en una posición con el decremento en efectividad según se aumente esta distancia. Por ejemplo, la palabra ***board*** en la colocación ***black board*** mantiene su sentido en todas las ocurrencias, independientemente del contexto en el que aparezca esta colocación. En (Martinez and Agirre, 2000) se desarrollaron distintos experimentos con palabras variando la sensibilidad de los sentidos (es decir, al utilizar palabras con sentidos bien diferenciados y otro conjunto con palabras con sentidos muy similares). Los resultados solamente empeoran cuando se consideran palabras con sentidos con diferencias sutiles.

2.5.4.4.3. UN SENTIDO POR DISCURSO

Introducida por (Gale *et al.*, 1992a) bajo el lema "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 en todo el discurso. Como fortaleza, se puede identificar que heurística cuando se aplica sobre palabras con sentidos bien diferenciados funciona bien. Por el contrario se obtiene malos. Esta idea se fundamenta con el estudio realizado por (Krovetz, 1998) donde demuestra que existen muchas palabras con sentidos similares y que además se utilizan juntas en el mismo discurso. Sin embargo luego de un análisis sobre un corpus, se demostró que el 70% de las palabras en este corpus tenía un solo sentido por discurso.

2.6. EVALUACIÓN DE LA DESAMBIGUACIÓN

Para evaluar en qué medidas los investigadores de la temática de WSD han progresado, se han desarrollado a lo largo de los años diversas competiciones. En estas los científicos exponen sus sistemas basados en técnicas de desambiguación y compiten. Los resultados obtenidos son posteriormente analizados por los respectivos jurados y revelan listados ordenados por puntuaciones, por lo general al medir Precisión (en inglés *Precision*) y Cobertura (en inglés *Recall*) según las ecuaciones (27) y (28) respectivamente.

La fórmula de precisión básicamente es la siguiente:

$$Precisión = \frac{sentidos\_correctos}{sentidos\_recuperados} \quad (27)$$

La fórmula de cobertura básicamente es:

$$Cobertura = \frac{sentidos\_correctos}{total\_sentidos} \quad (28)$$

$$F - Medida = \frac{2 * Precisión * Cobertura}{Precisión + Cobertura} \quad (29)$$

Otras de las medidas también utilizadas es la F-Medida (*F-Measure*), la cual establece un balance entre la precisión y la cobertura (véase a ecuación (29)). En las competiciones se exponen varias tareas en las que los sistemas compiten, estas son de diferentes temáticas del PLN. La competición que rige el *standing* de los sistemas de WSD se nombra competiciones de Senseval.

2.6.1. SENSEVAL. *EVALUATION EXERCISES FOR THE SEMANTIC ANALYSIS OF TEXT*

Debido a la necesidad de evaluar las diferentes aproximaciones para demostrar sus progresos en las tareas de PLN, fueron creadas la competiciones de Senseval[58] (*Evaluation Exercises for the Semantic Analysis of Text Senseval*). El primer Senseval (Senseval-1) se realizó en 1998 en Herstmonceux Castle, Succex en Inglaterra, y luego cada tres años una nueva competición tiene lugar. En Senseval, se definen diferentes tareas de PLN con el objetivo de evaluar sistemas a partir de repositorios y corpus comunes. Sin dicho marco común, sería muy difícil realizar comparaciones entre sistemas, además, la definición de diferentes tareas para evaluar la amplia variedad de sistemas proporciona como resultado la generación de recursos valiosos tanto para

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

el desarrollo como la evaluación de nuevos sistemas de WSD. Estas características han convertido a esta competición en un punto de partida para todos los investigadores, pues aquí se hacen públicos los resultados y se discuten las diferentes propuestas implicadas. Hasta el momento se han celebrado un total de cinco ediciones de estas competiciones. A continuación vamos a describir cada una de ellas, destacando aquellos sistemas más relevantes.

2.6.1.1. SENSEVAL-1. *FIRST EVALUATION EXERCISES FOR THE SEMANTIC ANALYSIS OF TEXT*

La primera edición de Senseval, constituyó un punto de partida, donde se establecieron las bases concretas del marco de evaluación. Estas bases implicaban repositorios de sentidos, posibles tareas a definir, conjuntos de evaluación, etc. En principio, se definió una única tarea relacionada con WSD, la llamada muestra léxica (en inglés, *Lexical Sample*). En esta tarea participaron un total de veinticinco sistemas, provenientes de veintitrés grupos de investigación diferentes. Los idiomas para los que se definió la tarea fueron: inglés, italiano y francés. Para el inglés en particular, se dispuso un conjunto de evaluación con treinta y cinco palabras con un total de 3.500 ocurrencias en diferentes contextos (Kilgariff and Palmer, 1998). Como inventario de sentidos se utilizó la base de datos léxica HÉCTOR (Atkins, 1993).

Los resultados reportados en Senseval-1 no emiten valores exactos, debido a ese detalle, los valores de los mejores sistemas se muestran según la interpretación aproximada de la gráfica de resultados de (Kilgarriff and Rosenzweig, 2000).

Entre los resultados más relevantes se obtuvo una exactitud aproximada al 74 % para el mejor de los sistemas, midiéndose con el *baseline* (*Most Frequent Sense* , MFS) que alcanzó un 57% (Kilgarriff and Rosenzweig, 2000). Además se puede decir que el comportamiento en general de los sistemas supervisados superó al resto de sistemas presentados en la competición.

Entre los mejores sistemas se encuentra el de (Hawkins and Nettleton, 2000) con un resultado muy próximo al 80% de *Precision* y *Recall*. Este sistema también es supervisado y combina diferentes técnicas: una técnica estocástica basada en la frecuencia de los sentidos en el texto, un conjunto de reglas extraídas de las asociaciones de palabras en el corpus de entrenamiento, y otra que trataba de obtener la similitud entre conceptos.

Otro sistema de relevantes resultados fue el sistema de Yarowsky con valores próximos al 80% de *Precision* y *Recall*, descrito con más detalle en (Yarowsky, 2000). Este sistema supervisado, estaba basado en listas de decisión jerárquicas, donde se introducían una serie de condiciones para la ramificación, reduciendo al mismo tiempo la excesiva fragmentación que los datos sufrían en los árboles de decisión. Esta aproximación utilizaba además varios tipos de atributos tales como: colocaciones, atributos morfológicos, sintácticos y contextuales, extraídos automáticamente de los datos de entrenamiento. A cada atributo se le asignaba un peso en función de la naturaleza del atributo y de su tipo.

Por otra parte, el sistema no supervisado que mejores resultados obtuvo fue *Suss*, desarrollado por (Ellman *et al.*, 2000). Este sistema obtuvo aproximadamente un 60% de exactitud en sus respuestas. Su estrategia fue la siguiente:

- Aplica un sistema básico que procesa los datos de entrenamiento (destacar que el sistema no se entrena, solamente se sirve de datos).
- Con un módulo de estadísticas, muestra la eficacia de desambiguación con la palabra, sentido de las palabras y el porcentaje de precisión.
- Luego, mide la eficacia sobre todo el corpus con las diferentes técnicas desarrolladas.
- Y por último utilización de técnicas que mejoran el rendimiento y eliminación de resultados más degradados.

2.6.1.2. SENSEVAL-2. *SECOND INTERNATIONAL WORKSHOP ON EVALUATING WORD SENSE DISAMBIGUATION SYSTEMS*

En Senseval-2 (*Second International Workshop on Evaluating Word Sense Disambiguation Systems*) se definieron tres tareas sobre doce idiomas y se presentaron alrededor de noventa sistemas. Esta competición tuvo lugar en Toulouse (Francia) en el año 2001. Una de las principales diferencias con la edición anterior, fue la creación de dos nuevas tareas: *All-Words*, cuyo objetivo es etiquetar con el sentido correcto todas las palabras con contenido semántico. Y una tarea similar a la tarea *Lexical Sample*, pero en la cual los sentidos se definieron a través de su traducción al japonés. En esta ocasión se empleó el repositorio de sentidos propuesto por WordNet 1.7, con el cual se anotaron los corpus de evaluación y *EuroWordNet* para el resto de idiomas.

Los sistemas participantes se distribuyeron dentro de las tareas de la siguiente forma:

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

Los resultados de las tareas de WSD se muestran en la Tabla 6 y Tabla 7. En general, los aciertos obtenidos en esta edición fueron peores que en la anterior edición. El mayor problema fue debido a que el repositorio de sentidos utilizado Senseval-2 era de mayor granularidad, consiguiendo aumentar la complejidad para resolver las ambigüedades. Nuevamente los mejores resultados fueron obtenidos por los sistemas supervisados en ambas tareas de WSD. Entre las técnicas que demostraron buenos resultados podemos destacar: votación de sistemas heterogéneos, uso de atributos complejos, selección de atributos y uso de material de entrenamiento adicional.

El mejor de los resultados en *Lexical Sample* lo obtuvo el sistema JHU (Yarowsky *et al.*, 2001), el cual empleaba una votación con varias configuraciones, validación cruzada de diferentes clasificadores supervisados (aplicando similitud de ejemplos, modelos Bayesianos y listas de decisión) y conjuntos de atributos complejos. Entre los atributos aplicados podemos destacar, las relaciones sintácticas o expresiones regulares construidas en base a las etiquetas morfológicas alrededor de la palabra objetivo.

El segundo puesto de esta competición en *Lexical Sample* lo lograron (Mihalcea and Moldovan, 2001) los cuales presentaron el sistema SMUls, además de contribuir en la tarea *All-Words* con el sistema SMU. En el caso de SMUls, se aplicó un algoritmo de aprendizaje basado en ejemplos, incluyendo una selección de atributos específicos por palabra, de tal modo que se seleccionaban aquellos atributos mejores para cada clasificador. Mientras que para la tarea *All-Words*, se empleó el mismo algoritmo cuando si se disponía de suficiente cantidad de ejemplos de entrenamiento para la palabra, en caso contrario se aplicó aprendizaje de patrones obtenidos desde el corpus SemCor, WordNet y otro corpus generado automáticamente. Dichos patrones se produjeron a partir del contexto local de las palabras y de cada *token*, al utilizar su forma base, su etiqueta morfológica, su sentido y su hiperónimo. Para el caso en el que tampoco se pudiera asignar el sentido mediante esta técnica, se asignaba el sentido de alguna ocurrencia de la misma palabra cercana en el texto y ya desambiguada, de la misma palabra cercana en el texto, o en última instancia del primer sentido de WN. Este sistema alcanzó el primer puesto en la tarea *All Words*, resultado que en ninguna competición hasta la fecha ha superado.

El sistema no supervisado que mejores resultados obtuvo en *All-Words* fue UNED (Fernández-Amorós *et al.*, 2009), este sistema aplicaba información mutua y la información de co-ocurrencia al mismo tiempo que algunas heurísticas de frecuencia para seleccionar un sentido. Basado en una matriz de relevancia, demostró ser ligeramente superior que el simple

conteo de co-ocurrencias. Lo que señala que de esta forma no se descartan palabras relevantes. En esta propuesta se identifica un problema que está asociado a las palabras que no resultan relevantes. El problema parece radicar en el hecho de que las palabras irrelevantes (con respecto a la palabra a desambiguar), rara vez se producen en el contexto de la palabra y en la definición de los sentidos (se analizan también las definiciones de los sentidos), siendo muy débil el impacto directo de la información en la matriz. Se puede decir entonces, que el sistema de SMU no es el único que usa definiciones de WordNet; sino que UNED también lo hace. Esto provoca que se marque entonces una pauta en el uso de conocimiento en sistemas de WSD.

Por lo general los sistemas que utilizan conocimiento semántico lo hacen basándose en clases semánticas obtenidas a partir del análisis de corpus (ej. sistema Sinequa-LIA-HMM), en definiciones de WordNet (ej. sistemas USM, ITT, DIMAP) y sus relaciones (ej. sistema Sheffield con el uso de distancias semánticas). Otro de los recursos semánticos utilizados en esta competición fue WND (por el sistema IRST). Se puede valorar entonces, que la participación con recursos semánticos es escasa y de uso independiente. Es importante destacar que el sistema IRST introdujo en esta competición la identificación de dominios asociados a las frases, idea que en futuras aproximaciones se defiende con nuevas propuestas.

English Lexical Sample - Fine-grained Scoring			
Precision	*Recall*	*System*	*Supervised*
0.642	0.642	*JHU (R)*	S
0.638	0.638	*SMUls*	S
0.629	0.629	*KUNLP*	S
0.617	0.617	*Stanford - CS224N*	S
0.613	0.613	*Sinequa-LIA - SCT*	S
0.594	0.594	*TALP*	S
0.571	0.571	*Duluth 3*	S
0.568	0.568	*JHU*	S
0.568	0.568	*UMD - SST*	S
0.573	0.564	*BCU - ehu-dlist-all*	S
0.554	0.554	*Duluth 5*	S
0.55	0.55	*Duluth C*	S
0.542	0.542	*Duluth 4*	S
0.539	0.539	*Duluth 2*	S
0.534	0.534	*Duluth 1*	S
0.523	0.523	*Duluth A*	S
0.512	0.512	*Baseline Lesk Corpus*	
0.508	0.508	*Duluth B*	S
0.498	0.498	*UNED - LS-T*	S
0.476	0.476	*Baseline Commonest*	
0.437	0.437	*Baseline Grouping Lesk Corpus*	
0.427	0.427	*Baseline Grouping Commonest*	
0.421	0.411	*Alicante*	S
0.402	0.401	*UNED - LS-U*	U
0.581	0.319	*ITRI - WASPS-Workbench*	U
0.293	0.293	*CL Research - DIMAP*	U
0.268	0.268	*Baseline Grouping Lesk*	
0.665	0.249	*IRST*	S
0.247	0.244	*IIT 2 (R)*	U
0.243	0.239	*IIT 1 (R)*	U
0.829	0.233	*BCU - ehu-dlist-best*	S
0.233	0.232	*IIT 2*	U
0.23	0.23	*Baseline Grouping Lesk Def*	
0.226	0.226	*Baseline Lesk*	
0.22	0.22	*IIT 1*	U
0.183	0.183	*Baseline Grouping Random*	
0.163	0.163	*Baseline Lesk Def*	
0.141	0.141	*Baseline Random*	

Tabla 6. Resultados de la tarea *Lexical Sample* para el inglés en Senseval-2.

English All Words - Fine-grained Scoring				
Precision	*Recall*		*System*	*Supervised*
0.690	0.690		*SMUaw*	S
0.669	0.646		*Baseline*-MFS-*Preiss*	-
0.636	0.636		*CNTS-Antwerp*	S
0.618	0.618		*Sinequa-LIA - HMM*	S
0.617	0.617		*Baseline*-MFS-*Chen*	
0.575	0.569		*UNED - AW-U2*	U
0.556	0.55		*UNED - AW-U*	U
0.475	0.454		*UCLA - gchao2*	S
0.474	0.453		*UCLA - gchao3*	S
0.416	0.451		*CL Research - DIMAP*	U
0.451	0.451		*CL Research - DIMAP* (*R*)	U
0.5	0.449		*UCLA - gchao*	S
0.36	0.36		*Universiti Sains Malaysia 2*	U
0.748	0.357		*IRST*	U
0.345	0.338		*Universiti Sains Malaysia 1*	U
0.336	0.336		*Universiti Sains Malaysia 3*	U
0.572	0.291		*BCU - ehu-dlist-all*	S
0.44	0.2		*Sheffield*	U
0.566	0.169		*Sussex - sel-ospd*	U
0.545	0.169		*Sussex - sel-ospd-ana*	U
0.598	0.14		*Sussex - sel*	U
0.328	0.038		*IIT 2*	U
0.294	0.034		*IIT 3*	U
0.287	0.033		*IIT 1*	U

Tabla 7. Resultados de la tarea *All Words* para el inglés en Senseval-2(los *baselines* son obtenidos de (Preiss, 2006) y (Chen *et al.*, 2010) respectivamente).

2.6.1.3. SENSEVAL-3. *THIRD INTERNATIONAL WORKSHOP ON THE EVALUATION OF SYSTEMS FOR THE SEMANTIC ANALYSIS OF TEXT*

La tercera competición de Senseval tuvo lugar en julio del 2004 en Barcelona (España). En esta edición se organizaron catorce tareas (más información en www.senseval.org/senseval3). De ellas, dos fueron sobre desambiguación *All-Words* y *Lexical Sample*, la primera incluyendo idiomas como inglés (64 sistemas) e italiano (7 sistemas). *Lexical sample* se aplicó para inglés (65 sistemas), italiano (11 sistemas), vasco (8 sistemas), catalán (8 sistemas), chino (16 sistemas), rumano (8 sistemas), español (18 sistemas) y multilingüe (23 sistemas) (los textos originales eran en inglés y la anotación para las palabras se debía hacer para la traducción en otro idioma (a textos ingles-francés e inglés-hindi)) y otras tareas como *Automatic subcategorization acquisition* (35 sistemas) (en esta tarea se evaluaron diversos sistemas de WSD en el contexto de sub-categorización automática), WSD *of WordNet glosses* (36 sistemas) (el objetivo de esta tarea era desarrollar un método de anotación automática tomando como corpus de evaluación las glosas previamente etiquetadas en *eXtended* WordNet), *Semantic Roles* (36 sistemas) (utiliza 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 and Jurafsky, 2002), y *Logic Forms* (26 sistemas) (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 corresponde con un predicado). Las versiones de WordNet utilizadas en estas tareas fueron desde 1.6 hasta 1.7.1.

A continuación se relatan únicamente la Tarea 1 (*English All Words*) y la Tarea 6 (*English Lexical Sample*) por ser las más acordes con el objetivo de la Tesis.

- Tarea 1 (Snyder and Palmer, 2004). Tal y como se realizó en Senseval-2, en esta nueva edición se etiquetaron aproximadamente cinco mil palabras extraídas del corpus de

Penn Treebank[59] tomando como inventario de sentidos WordNet 1.7.1. Se etiquetaron nombres, adjetivos y adverbios haciéndose dos revisiones para formalizar criterios.

- Tarea 6. (Mihalcea *et al.*, 2004) Los datos en esta tarea se obtuvieron a partir de la interfaz del *Open Mind Word Expert* (OMWE) (Chklovski and Mihalcea, 2002). Para asegurar la fiabilidad, se extrajeron dos etiquetas por elemento y se realizaron diversas pruebas con la finalidad de llegar a un acuerdo entre las distintas anotaciones de los etiquetadores. Se extrajeron alrededor de sesenta palabras ambiguas entre nombres, adjetivos y verbos. Parte de las pruebas de evaluación fueron creadas por el Departamento de Lingüística de la Universidad del Norte de Texas (UNT). Otra parte fue extraída a partir de corpus etiquetado de la web. De igual manera el inventario de sentidos fue WordNet 1.7.1 para nombres y adjetivos y *Wordsmyth-2*[60] para verbos. Para no hacer tan fina la sensibilidad de los sentidos, se aplicaron dos modos de evaluación, gruesa (*coarse*) donde los sentidos considerados muy próximos entre sí podían agruparse y tomarse como un único sentido y una más fina (*fine*) donde la frontera entre un sentido u otro estaba muy ajustada.

En las tareas de desambiguación descritas, los mejores sistemas fueron nuevamente los supervisados. En *Lexical Sample* el sistema *htsa3* (Grozea, 2004) obtuvo los mejores resultados. Este utiliza *Regularized Least Square Classification* (RLSC) (Popescu, 2004) para la extracción de características que sirven de parámetros en un sistema de aprendizaje Bayesiano. El entrenamiento incorpora el uso de la corrección de las frecuencias, al dividir la confianza de salida de los sentidos por la frecuencia$^{\alpha}$ (α=0,2).

El segundo mejor sistema es el IRST-Kernels (Strapparava *et al.*, 2004). Este sistema implementa una aproximación basada en una función *kernel* que combinaba diferentes fuentes de información. Se consideran entonces dos funciones, una sintagmática y otra paradigmática. En la sintagmática se modela la similitud de dos contextos (secuencias de palabras comunes), donde se consideran colocaciones de palabras y etiquetas morfológicas. En la paradigmática se extrajen los dominios asociados al texto. En esta segunda función intervienen dos *kernels*, uno consistente en una bolsa de palabras alrededor de la palabra objetivo, y otro que implementa la técnica LSA .

En la tarea *All Words* el mejor sistema fue el GAMBL (Decadt *et al.*, 2004). Este sistema se basa en la idea de utilizar un sistema de clasificación para resolver la ambigüedad. Cada clasificador se especializa en una palabra. Para entrenar el sistema de expertos se utilizó un sistema de aprendizaje basado en memoria (*Memory Based Learning*, MBL). La función conjunta de selección y optimización de parámetros del algoritmo se logran con un algoritmo genético. Además, tiene en cuenta un enfoque de clasificación en cascada, en la que el algoritmo genético optimiza las características de contexto local y la salida de un clasificador de palabras clave. A diferencia de las versiones anteriores basadas en memoria, se añade el uso de relaciones gramaticales y las características de estas. El entrenamiento del sistema recibe información proveniente de ejemplos como SemCor, corpus de Senseval anteriores, WordNet y otros. Los atributos extraídos se enfocan en el análisis del contexto local, y de un conjunto de palabras claves que representan a las palabras objetivo.

El segundo mejor sistema fue *SenseLearner* (Mihalcea and Faruque, 2004). Sistema mínimamente supervisado, el cual usa ejemplos etiquetados manualmente, además de SemCor y WordNet. En una primera fase (Modelo de Lenguaje Semántico), utiliza un conjunto de modelos aprendidos mediante una técnica de aprendizaje automático basada en ejemplos del corpus de SemCor. En una segunda fase (Generalización Semántica al usar Dependencias

[59] http://www.cis.upenn.edu/~treebank/
[60] http://www.wordsmyth.net/

Sintáctica y una Red Conceptual), por medio de WordNet y SemCor, se extraen dependencias entre sustantivos y verbos, para luego abstraerse hacia sus hiperónimos para generalizar el patrón de relación (ej. *drink water* sus hiperónimos serían *take-in liquid*) . Posteriormente, al emplear un algoritmo de aprendizaje basado en estos ejemplos se genera un clasificador capaz de desambiguar las palabras objetivo.

De los sistemas sin supervisión con mejores resultados se puede destacar IRST-DDD (Strapparava *et al.*, 2004). Este sistema aplicaba el mismo principio que el empleado en Senseval-2, pero agregando algunas variantes diferentes. En esta ocasión se obtienen dos tipos de vectores de dominios, (I) vectores de *synset*, donde se considera la relevancia de un *synset* respecto a cada dominio, y (II) vectores del texto, donde se considera la relevancia de cada porción del texto con respecto al conjunto. El núcleo de IRST-DDD se basa en establecer un *ranking* de comparación entre los vectores generados por cada *synset* con respecto a los vectores que genera el contexto. Para la construcción de los vectores de los *synsets* se utiliza el recurso WND, mientras que el cálculo de los valores de comparación se analiza las probabilidades del *synset* en SemCor. Otra variante de IRST aplica LSA, donde aumenta la información de los dominios asociados a los sentidos, de este modo se enriquece el conocimiento obtenido anteriormente. Esta segunda propuesta aunque es interesante no logra superar los resultados de la primera.

En esta edición de Senseval se identificaron varios sistemas basados en conocimiento. Por ejemplo, el que se acaba de describir (IRST) y DLSI-UA (Vázquez *et al.*, 2004b), utilizan los dominios de (Magnini and Cavaglia, 2000) mediante la detección de Dominios Relevantes en textos. Otro sistema, *Meaning* (Villarejo *et al.*, 2004), introduce como parámetros de entrenamiento, las características sintácticas del texto, y semántica mediante la exploración del recurso semántico MRC (Atserias *et al.*, 2004).

En la Tabla 8 y Tabla 9 se pueden observar los resultados obtenidos por los sistemas en ambas tareas, donde se aprecia un comportamiento similar a la competición anterior entre sistemas supervisados y no supervisados. Es importante destacar que los mejores fueron inferiores en comparación con la competición anterior (Senseval-2).

System/Team	***Fine* (%)**		***Coarse* (%)**		***Supervised***
	P	R	P	R	
htsa3 \ U.Bucharest (*Grozea*)	72.9	72.9	79.3	79.3	S
IRST-Kernels \ ITC-IRST (*Strapparava*)	72.6	72.6	79.5	79.5	S
nusels \ Nat.U. Singapore (*Lee*)	72.4	72.4	78.8	78.8	S
htsa4	72.4	72.4	78.8	78.8	S
BCU comb \ Basque Country U. -(*Agirre & Martínez*)	72.3	72.3	78.9	78.9	S
htsa1	72.2	72.2	78.7	78.7	S
rlsc-comb \ U.Bucharest (*Popescu*)	72.2	72.2	78.4	78.4	S
htsa2	72.1	72.1	78.6	78.6	S
BCU english	72.0	72.0	79.1	79.1	S
rlsc-lin	71.8	71.8	78.4	78.4	S
HLTC HKUST all \ HKUST (*Carpuat*)	71.4	71.4	78.6	78.6	S
TALP\U.P.Catalunya - (*Escudero et al.*)	71.3	71.3	78.2	78.2	S
MC-WSD \ Brown U. - (*Ciaramita & Johnson*)	71.1	71.1	78.1	78.1	S
HLTC HKUST all2	70.9	70.9	78.1	78.1	S
NRC-Fine \ NRC (*Turney*)	69.4	69.4	75.9	75.9	S
HLTC HKUST me	69.3	69.3	76.4	76.4	S
NRC-Fine2	69.1	69.1	75.6	75.6	S
GAMBL \U. Antwerp (*Decadt*)	67.4	67.4	74.0	74.0	S
SinequaLex \ Sinequa Labs (*Crestan*)	67.2	67.2	74.2	74.2	S
CLaC1 \ Concordia U. (*Lamjiri*)	67.2	67.2	75.1	75.1	S
SinequaLex2	66.8	66.8	73.6	73.6	S
UMD SST4 \ U. Maryland (*Cabezas*)	66.0	66.0	73.7	73.7	S
wsdiit \ IIT Bombay - (*Ramakrishnan et al.*)	66.1	65.7	73.9	74.1	U
Prob1 \ Cambridge U. (*Preiss*)	65.1	65.1	71.6	71.6	S
SyntaLex-3 \ U.Toronto (*Mohammad*)	64.6	64.6	72.0	72.0	S

UNED (*Artiles*)	64.1	64.1	72.0	72.0	S
SyntaLex-4	63.3	63.3	71.1	71.1	S
CLaC2	63.1	63.1	70.3	70.3	S
SyntaLex-1	62.4	62.4	69.1	69.1	S
Prob2	61.9	61.9	69.3	69.3	S
SyntaLex-2	61.8	61.8	68.4	68.4	S
Duluth-ELSS \ U.Minnesota (*Pedersen*)	61.8	61.8	70.1	70.1	S
UJAEN \ U.Jáen (*García-Vega*)	61.3	61.3	69.5	69.5	S
Cymfony (*Niu*)	56.3	56.3	66.4	66.4	U
MFS *Baseline*	55.2	55.2	64.5	64.5	
Prob0 \ Cambridge U. (*Preiss*)	54.7	54.7	63.6	63.6	U
R2D2 \ U. Alicante (*Vázquez*)	63.4	52.1	69.7	57.3	S
IRST-Ties \ ITC-IRST (*Strapparava*)	70.6	50.5	76.7	54.8	S
NRC-Coarse	48.5	48.5	75.8	75.8	S
NRC-Coarse2	48.4	48.4	75.7	75.7	S
clr04-ls \ CL Research - (*Litkowski*)	45.0	45.0	55.5	55.5	U
CIAOSENSO \ U. Genova (*Buscaldi*)	50.1	41.7	59.1	49.3	U
KUNLP \ Korea U. (*Seo*)	40.4	40.4	52.8	52.8	U
Duluth-SenseRelate \ U.Minnesota (*Pedersen*)	40.3	38.5	51.0	48.7	U
DLSI-UA-LS-SU \ U.Alicante (*Vázquez*)	78.2	31.0	82.8	32.9	S
DFA-LS-Unsup \ UNED (*Fernández*)	23.4	23.4	27.4	27.4	U
DLSI-UA-LS-NOSU \ U.Alicante (*Vázquez*)	19.7	11.7	32.2	19.0	U

Tabla 8. *Ranking* de la competición Senseval-3, Tarea 6 (*Precision* (P), *Recall* (R), *Supervised* (S) (Supervisado), *Unsupervised* (U) (Sin Supervisión))).

System	**P (%)**	**R (%)**	***Supervised***
GAMBL-AW	65.1	65.1	S
Sense-Learner	65.1	64.2	S
Koc University	64.8	63.9	S
R2D2 English-All-Word	62.6	62.6	-
MFS *Baseline* (*GAMBL-AW*)	62.4	62.4	
Meaning All-words	62.5	62.3	S
Meaning simple	61.1	61	S
MFS *Baseline* (*Yuret*)	60.9	60.9	
LCCaw	61.4	60.6	-
upv-shmm-eaw	61.6	60.5	-
UJAEN	60.1	58.8	S
IRST-DDD-00	58.3	58.2	U
Sussex-Prob5	58.5	56.8	-
Sussex-Prob4	57.5	55	-
Sussex-Prob3	57.3	54.7	-
DFA-Unsup-AW	55.7	54.6	U
KUNLP-Eng-All	51	49.6	U
IRST-DDD-LSI	66.1	49.6	U
upv-unige-CIAOSENSO-eaw	58.1	48	U
merl.system3	46.7	45.6	-
upv-unige-CIAOSENSO3-eaw	60.8	45.1	U
merl-system1	45.9	44.7	-
IRST-DDD-09	72.9	44.1	U
autoPS	49	43.3	U
clr-04-aw	50.6	43.1	-
autoPSNVs	56.3	35.4	U
merl.system2	48	35.2	-
DLSI-UA-All-Nosu	34.3	27.5	-

Tabla 9. *Ranking* de la competición Senseval-3, Tarea 1 (*Precision* (P), *Recall* (R), *Supervised* (S)(Supervisado), *Unsupervised* (U) (Sin Supervisión).

2.6.1.4. SEMEVAL-1. *FOURTH INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATIONS*

La competición Semeval-1 tuvo lugar en junio del 2007 en Praga (República Checa). En esta edición se organizaron diecinueve tareas cancelándose la Tarea 3, compitiéndose entonces en dieciocho de estas. En esta edición se cambia el nombre de *SensEval* a *SemEval*, del inglés *Semantic Evaluations* (*Agirre et al., 2007*), debido a que se incluyen numerosas tareas relacionadas con el análisis semántico de textos en general, y no solamente con variantes de WSD.

- Tarea 1. *Evaluating WSD on Cross-Language Information Retrieval.* (2 sistemas)
- Tarea 2. *Evaluating Word Sense Induction and Discrimination Systems.* (6 sistemas)
- Tarea 3. (***Canceleda***)*: Pronominal Anaphora Resolution in the Prague Dependency Treebank 2.0*
- Tarea 4. *Classification of Semantic Relations between Nominals.* (15 sistemas)
- Tarea 5. *Multilingual Chinese-English Lexical Sample.* (6 sistemas)
- Tarea 6. *Word Sense Disambiguation of Prepositions.* (5 sistemas)
- Tarea 7. *Coarse Grained English All Words Task.* (12 sistemas)
- Tarea 8. *Metonymy Resolution at SemEval 2007.* (5 sistemas)
- Tarea 9. *Multilevel Semantic Annotation of Catalan and Spanish.* (*2 sistemas*)
- Tarea 10. *English Lexical Substitution Task.*
- Tarea 11. *English Lexical Sample Task via English-Chinese Parallel Text.*
- Tarea 12. *Turkish Lexical Sample Task.*
- Tarea 13. *Web People Search.*
- Tarea 14. *Affective Text.*
- Tarea 15. *TempEval Temporal Relation Identification.*
- Tarea 16. *Evaluation of Wide Coverage Knowledge Resources.*
- Tarea 17. *English Lexical Sample, SRL and All Words.* (13 sistemas en *Lexical Sample*, 15 sistemas en *All Words*)
- Tarea 18. *Arabic Semantic Labeling.*
- Tarea 19. *Frame Semantic Structure Extraction.*

Resulta importante resaltar que en esta edición de Senseval, se tuvo en cuenta un tema debatido en Senseval-3, acerca de la problemática generada por la granularidad de los sentidos del inventario utilizado. En concreto, que el uso de los sentidos demasiado detallados de WN no resulta de mucha utilidad para otras aplicaciones de más alto nivel dentro del campo del PLN. Por ese motivo, se reconoce que al hacer uso de este repositorio de sentidos de alta granularidad, es muy difícil superar los resultados obtenidos. Considerando estas valoraciones, se propuso agrupar algunos sentidos de mayor sensibilidad con el objetivo de generar otro repositorio de menor granularidad y utilizarlo en las tareas tradicionales. Esta idea demostró un incremento considerable en las exactitudes de WSD en la Tarea 7, con resultados para el análisis de todas la palabras (*All-Words*) con valores alrededor del 82% (Navigli *et al.*, 2007) y Tarea 17 (*lexical simple*)del 88% de F-Medidas (balance entre Precisión y Cobertura).

Para la tarea *Lexical Sample* en el análisis de los sentidos con granularidad gruesa, se tomó como repositorio de sentidos el recurso *Ontonotes*[61]. En esta tarea participaron trece sistemas y se incluyeron dos sistemas más que se notificaron tarde. La desambiguación se le aplicó a 4851 ocurrencias de cien palabras distintas (Pradhan *et al.*, 2007). Entre todos los resultados se evidencia exactitudes muy superiores a los obtenidos en la misma tarea en Senseval-3. Esto es

[61] http://www.bbn.com/ontonotes

debido a la reducción de la granularidad. Tal y como se muestra en la Tabla 10 el mejor de los sistemas fue el NUS-ML (*Cai et al., 2007*) con los mismos valores de precisión y cobertura. En este sistema se aplican clasificadores de aprendizaje automático combinando modelos bayesianos estructurados en una jerarquía de tres niveles. Los atributos del entrenamiento son léxicos, sintácticos y atributos de tópico. El segundo mejor resultado lo obtuvo el sistema UBC-ALM (*Agirre and de_Lacalle, 2007*) también con los mismos valores de precisión y cobertura. Esta propuesta combina diferentes clasificadores K-NN basados en similitud de ejemplos, donde se utilizan atributos locales, de tópico y latentes. Estos últimos son obtenidos mediante una técnica Descomposición de Valores Singulares (SVD).

En cuanto a la tarea de evaluación de todas las palabras (*All Words*)(véase la Tabla 11), se propusieron dos de estas tareas, la tradicional haciendo uso de sentidos finos de WordNet (Tarea 17), y una nueva tarea con el empleo de sentidos más gruesos (Tarea 7). En este caso, se analiza la Tarea 17, por ser la que necesita mayor mejora en los resultados obtenidos.

El mejor de los sistemas en *All Words* para sentidos finos fue un sistema de aprendizaje automático basado en las técnicas de máxima entropía. PNNL (Tratz *et al.*, 2007) extrae tres tipos de características de los textos, información contextual (con el análisis de tres *tokens* alrededor de la palabra objetivo sin límites en el texto), información sintáctica (incluyendo dependencias gramaticales (ej. sujeto-objeto), características sintáctico morfológicas (ej. tiempo, número y categoría gramatical)) e información semántica (tipos de entidades nombradas (ej. Persona, Localización, Organización) e hiperonimia (obtención del ancestro)).

El segundo mejor sistema fue NUS-PT (Chan *et al.*, 2007) también supervisado, con el uso de SVM (*Support Vector Machines*). Las fuentes de conocimiento que tiene en consideración incluyen la colocación local de las categorías gramaticales y palabras acompañantes. El aprendizaje de este sistema se aplica sobre ejemplos de corpus paralelos del inglés-chino, SemCor y DSO[62] además del inventario de WordNet.

El mejor de los sistemas no supervisados fue el sistema RACAI (Ion and Tufis, 2007). Este sistema se basa en la suposición de que el mejor candidato de los significados de las palabras a ser asignado, es el que se rige por la construcción de una interpretación de la frase completa. Dicha interpretación se ve facilitada por la especificación del contexto en dependencia de una palabra dentro de la oración. Esto significa, que se rige por realizar un análisis de dependencia pseudo-sintáctico de la oración y determinar de esta manera la definición de la palabra objetivo más apropiada.

En la Tabla 11 se muestran los resultados obtenidos por los sistemas mencionados, además de las técnicas aplicadas por cada sistema en la competición. Se aprecian pocos sistemas que se sirven de conocimientos externos con la excepción de WordNet. Por ejemplo, UPV-WSD (Buscaldi and Rosso, 2007) con un resultado significativo entre los sistemas no supervisados, aplica técnicas difusas y el cálculo de Densidad Conceptual descrito en la sección 2.5.4.2.1. Dicho sistema utiliza el recurso de WND para el cálculo de la medida. JU-SKNSB (Naskar and Bandyopadhyay, 2007) y UBC-UMB (Martinez et al., 2007) utiliza el recurso *eXtended* WordNet, el segundo de estos sistemas se sirve además de las relaciones internas de MCR (*Multilingual Central Repository*).

[62] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12

Rank	*Participant*	*System*	*Classifier*	F (%)
1	Cai Junfu	NUS-ML	SVM	88.7±1.2
2	Oier López de Lacalle	UBC-ALM	SVD+kNN	86.9±1.2
3	Zheng-Yu Niu	I2R	*Supervised*	86.4±1.2
4	Lucia Specia	USP-IBM-2	SVM	85.7±1.2
5	Lucia Specia	USP-IBM-1	ILP	85.1±1.2
5	Deniz Yuret	KU	*Semi-supervised*	85.1±1.2
6	Saarikoski	OE	Naive Bayes, SVM	83.8±1.2
7	University of Technology Brno	VUTBR	Naive Bayes	80.3±1.2
8	Ana Zelaia	UBC-ZAS	SVD+kNN	79.9±1.2
9	Carlo Strapparava	ITC-iRST	SVM	79.6±1.2
10	*MFS Baseline*	*Baseline*	N/A	78.0±1.2
11	Toby Hawker	USYD	SVM	74.3±1.2
12	Siddharth Patwardhan	UMND1	*Unsupervised*	53.8±1.2
13	Saif Mohammad	Tor	*Unsupervised*	52.1±1.2
-	Toby Hawker	USYD	SVM	89.1±1.2
-	Carlo Strapparava	ITC	SVM	89.1±1.2

Tabla 10. *Ranking* de resultados de la competición Semeval-1 para la tarea 17 en *Lexical Sample*, F(F-*Measure* (F-Medida)) (Pradhan *et al.*, 2007).

Rank	*Participant*	*System*	*Classifier*	F (%)
1	Stephen Tratz	PNNL	MaxEnt	59.1±4.5
2	Hwee Tou Ng	NUS-PT	SVM	58.7±4.5
3	Rada Mihalcea	UNT-Yahoo	*Memory-based*	58.3±4.5
4	Cai Junfu	NUS-ML	Naive Bayes	57.6±4.5
5	Oier López de Lacalle	UBC-ALM	kNN	54.4±4.5
6	David Martínez	UBC-UMB-2	kNN	54.0±4.5
7	Jonathan Chang	PU-BCD	*Exponential Model*	53.9±4.5
8	Radu ION	RACAI	*Unsupervised*	52.7±4.5
9	*MFS WordNet*	*Baseline*	N/A	51.4±4.5
10	Davide Buscaldi	UPV-WSD	*Unsupervised*	46.9±4.5
11	Sudip Kumar Naskar	JU-SKNSB	*Unsupervised*	40.2±4.5
12	David Martínez	UBC-UMB-1	*Unsupervised*	39.9±4.5
14	Rafael Berlanga	tkb-uo	*Unsupervised*	32.5±4.5
15	JordanBoyd-Graber	PUTOP	*Unsupervised*	13.2±4.5

Tabla 11. *Ranking* de resultados de la competición Semeval-1 para la tarea 17 en *English All-Word*, *F* (F-*Measure* (F-Medida)).

2.6.1.5. SEMEVAL-2. *5TH INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION*

Semeval-2 ha sido la quinta edición de las competiciones, la cual tuvo lugar en Uppsala (Suecia) en julio del 2010. Esta edición coincidió con la conferencia internacional ACL (*Association for Computational Linguistics*). En SemEval-2 se propusieron un total de 18 tareas, de naturaleza muy variada. Respetando el motivo de su creación, se presentaron tareas relacionadas con la desambiguación semántica, pero además se añadieron nuevas tareas. A continuación se mencionan las tareas presentadas en esta edición:

- Tarea 1. *Coreference Resolution in Multiple Languages*
- Tarea 2. *Cross-Lingual Lexical Substitution* (14 sistemas)
- Tarea 3. *Cross-Lingual Word Sense Disambiguation* (7 sistemas en cada idioma, español y francés, y 3 sistemas en cada idioma, el italiano, danés y alemán)
- Tarea 4. *VP Ellipsis - Detection and Resolution*
- Tarea 5. *Automatic Keyphrase Extraction from Scientific Articles* (19 sistemas)
- Tarea 6 (**Cancelada**). *Classification of Semantic Relations between MeSH Entities in Swedish Medical Texts*
- Tarea 7. *Argument Selection and Coercion*

- Tarea 8. *Multi-Way Classification of Semantic Relations Between Pairs of Nominals* (28 sistemas)
- Tarea 9. *Noun Compound Interpretation Using Paraphrasing Verbs* (*7 sistemas*)
- Tarea 10. *Linking Events and their Participants in Discourse* (2 sistemas en cada subtarea)
- Tarea 11. *Event Detection in Chinese News Sentences* (7 sistemas)
- Tarea 12. *Parser Training and Evaluation using Textual Entailment* (19 sistemas)
- Tarea 13. *TempEval 2*
- Tarea 14. *Word Sense Induction* (*28 sistemas incluyendo los baselines* MFS *y Random*)
- Tarea 15. *Infrequent Sense Identification for Mandarin Text to Speech Systems* (9 sistemas)
- Tarea 16. *Japanese* WSD
- Tarea 17. *All-words Word Sense Disambiguation on a Specific Domain* (WSD-*domain*) (29 sistemas en el inglés, 5 en el chino, 3 en el danés y 3 en el italiano)
- Tarea 18. *Disambiguating Sentiment Ambiguous Adjectives* (16 sistemas)

La tradicional tarea *All Words*, se aplica ahora para un dominio restringido, se enfoca en el estudio de la dependencia del dominio de los sistemas de WSD. Para ello, los textos de evaluación de los sistemas pertenecen a dominios específicos, relacionados con el medio ambiente. Por esta vía se pudo comparar los sistemas que entrenaban con corpus de propósito general con los que utilizaron corpus especializados, cuando el objetivo es texto relativo a diferentes dominios. En esta tarea la aproximación presentada en el apartado 4.2.1 de esta Tesis forma parte de los sistemas participantes.

Como se aprecia en la Tabla 12 los mejores sistemas nuevamente son los que aplican técnicas de supervisión. El mejor de los sistemas para la desambiguación de todas las palabras del idioma inglés es CFILT (Kulkarni *et al.*, 2010), el cual se presenta con tres versiones en dos variantes: con técnicas débilmente supervisadas y solamente basadas en conocimiento. La primera versión obtuvo el mejor de los resultados del *ranking*, con la selección de "Un Sentido Por Dominio". En ella se aplica un primer paso de supervisión con el objetivo de detectar los sentidos más frecuentes de las palabras implicadas. Luego, se pasa a la ejecución de la propuesta de WSD interactiva de (Khapra *et al.*, 2010), donde originalmente se introducen sentidos monosémicos como semillas, para generar una cadena de ejecuciones alrededor de estos e ir identificando los nuevos sentidos correctos. En esta ocasión se introducen como semillas los sentidos más frecuentes. En su otra versión (el sistema basado en conocimiento), se genera un grafo a partir de la representación de sentidos de las palabras que se encuentran en corpus no anotados, mediante el uso de WordNet, propiciando la exploración de las interrelaciones existentes. A partir de este tipo de base de conocimiento generada, se pueden identificar palabras coexistentes en un mismo corpus de dominio. Ambas propuestas ocupan las primeras posiciones del *ranking* en su tipo (ej. WS y KB, véase la Tabla 12).

El segundo mejor sistema fue IIITH (Reddy et al., 2010), donde también se aplican técnicas débilmente supervisadas con la ayuda de informaciones resultantes de algoritmos basados en grafos, concretamente se utiliza el *Personalized PageRank* (Agirre and Soroa, 2009). En la ejecución del sistema IIITH utiliza como base de conocimiento WordNet 3.0 + Glosas desambiguadas (conocido como *Lexical Knowledge Base* (LKB) de WordNet). Con el uso del *PageRank* es posible ponderar ciertas relaciones entre nodos y diferenciar así vínculos de mayor y menor importancia. En este sistema en particular se aplican para establecer pesos iniciales las medidas de Densidad Conceptual descritas en la sección 2.5.4.2.1.

La segunda mejor propuesta basada en conocimiento es *TreeMatch* (Tran *et al.*, 2010). En este sistema, se utilizan las glosas de los *synsets* de WordNet y se realiza un *matching* entre glosas y la frase original. El objetivo es seleccionar el sentido de mayor solapamiento, utilizando árboles de dependencia extraídos del texto original y cada definición (glosa).

También se puede destacar del sistema Kytoto (Soroa *et al.*, 2010), el cual aplica la base de conocimientos LKB de UKB (sistemas propuesto por (Agirre and Soroa, 2009)) e igualmente aplica el conocido algoritmo de *PageRank* introducido por primera vez por (Brin and Page, 1998).

Rank	*System*	*Type*	*Precision*	*Recall*
1	CFILT-2	WS	0.570	0.555
2	CFILT-1	WS	0.554	0.540
3	IIITH1-d.l.ppr.05	WS	0.534	0.528
4	IIITH2-d.r.l.ppr.05	WS	0.522	0.516
5	BLC20*SemcorBackground*	S	0.513	0.513
-	MFS *baseline*	-	0.505	0.505
6	BLC20Semcor	S	0.505	0.505
7	CFILT-3	KB	0.512	0.495
8	Treematch	KB	0.506	0.493
9	Treematch-2	KB	0.504	0.491
10	kyoto-2	KB	0.481	0.481
11	Treematch-3	KB	0.492	0.479
12	RACAI-MFS	KB	0.461	0.460
13	UCF-WS	KB	0.447	0.441
14	HIT-CIR-DMFS-1.ans	KB	0.436	0.435
15	UCF-WS-domain	KB	0.440	0.434
16	IIITH2-d.r.l.*baseline*.05	KB	0.496	0.433
17	IIITH1-d.l.*baseline*.05	KB	0.498	0.432
18	RACAI-2MFS	KB	0.433	0.431
19	IIITH1-d.l.ppv.05	KB	0.426	0.425
20	IIITH2-d.r.l.ppv.05	KB	0.424	0.422
21	UCF-WS-*domain.noPropers*	KB	0.437	0.392
22	kyoto-1	KB	0.384	0.384
23	BLC20Background	S	0.380	0.380
24	NLEL-WSD-PDB	WS	0.381	0.356
25	RACAI-Lexical-Chains	KB	0.351	0.350
26	NLEL-WSD	WS	0.370	0.345
27	*Relevant Semantic Trees*	KB	0.328	0.322
28	*Relevant Semantic Trees-2*	KB	0.321	0.315
29	*Relevant Cliques*	KB	0.312	0.303
-	*Random baseline*	-	0.23	0.23

Tabla 12. *Ranking* de sistemas para la tarea *All-words Word Sense Disambiguation on a Specific Domain* del inglés en Semeval-2. (Sistema débilmente supervisado (*weak Supervised* (WS)), supervisado (*Supervised* (S)) y basado en conocimiento (*knowlodge-based* (KB))).

En esta última competición de Senseval se hace patente que los algoritmos basados en grafos demuestran ser líderes en ante los nuevos retos que ofrece WSD. A continuación se muestra en la Tabla 13 una comparativa de las diferentes ediciones de la competición Senseval. Los datos corresponden con la tarea *All Words* con la excepción de la primera de las competiciones, donde no se compite en esa tarea y por tanto se ha excluido de la tabla. Lo más importante a observar, es el tipo de sistema que lideró cada competición y el comportamiento del *baseline* MFS.

Sistema	**Senseval-2 (2001)**			
	P(%)	R(%)	Tipo	Técnica
Líder total	69	69	S	Aprendizaje de patrones
Líder sin supervisión	57.5	56.9	U/KB	Información Mutua
MFS	66.9	64.6	-	
	61.7	61.7		
Sistema	Senseval-3 (2004)			
	P(%)	R(%)	Tipo	Técnica
Líder total	65.1	65.1	S	Aprendizaje basado en memoria
Líder sin supervisión	58.3	58.2	U/KB	Dominios Relevantes
MFS	62.4	62.4	-	
	60.9	60.9		
Sistema	Semeval-1(2007)			
	P(%)	R(%)	Tipo	Técnica
Líder total	59.1	59.1	S	Máxima entropía
Líder sin supervisión	52.7	52.7	U	Dependencia pseudo-sintáctico
MFS	51.4	51.4	-	
Sistema	Semeval-2(2010)			
	P(%)	R(%)	Tipo	Técnica
Líder total	57	55.5	WS	MFS, WSD interactiva/Basado en Grafos
Líder sin supervisión	51.2	49.5	U/KB	Basado en grafos
MFS	50.5	50.5	-	

Tabla 13. Comparativa entre competiciones de Senseval (*Precision* (P), *Recall* (R), Supervisado(S), Sin Supervisión (U), Basado en conocimiento (KB)).

Como se puede apreciar con el paso de los años, los resultados de los sistemas de WSD empeoran, lo que no quiere decir que las propuestas sean peores, sino que los corpus de evaluación son cada vez más exigentes. Es obvio que las aproximaciones más exactas son las que aplican algún tipo de aprendizaje, pero con el paso del tiempo la introducción de técnicas basadas en conocimiento y exploración de sus estructuras, han recortado la diferencia entre las puntuaciones obtenidas por sistemas con algún tipo de supervisión y los que no. Este progreso es sumamente importante, debido a que los sistemas basados en conocimiento son indiferentes al entorno del corpus a ser desambiguado y no requieren de grandes cantidades de texto anotado manualmente para ejecutarse.

2.7. CONCLUSIONES

Este capítulo se ha estudiado la temática asociada a la desambiguación del sentido de las palabras y los trabajos precedentes. Este estudio ha sido de vital importancia para constituir las bases de la investigación, pues se han descrito todos los componentes esenciales asociados a la tarea de desambiguación del sentido de las palabras. Se comenzado por describir tres elementos fundamentales que intervienen en el proceso de WSD. El primero hace referencia al proceso de selección de sentidos correctos. Seguido se ha abordado cómo se representa el contexto de la palabra y por último las fuentes de conocimiento externas de las que se sirven los sistemas de desambiguación (corpus etiquetados, los no etiquetados, los recursos semánticos y otros). Entre estas fuentes de conocimiento, figura uno de los recursos más difundidos en la actualidad (WordNet), que a su vez ha servido como inspiración en la creación de muchos otros. Los recursos léxicos y semánticos en general, han sido útiles en muchas áreas de PLN, lo que ha motivado a su integración como una única herramienta, obteniendo como resultado bases de conocimiento multilingües y semánticamente multidimensionales.

Con el objetivo de conocer el desarrollo de las propuestas de tratamiento semántico, se han analizado varios métodos básicos de WSD agrupados según su clasificación. Por ejemplo, los métodos supervisados requieren de grandes cantidades de corpus etiquetados manualmente para

que sus sistemas aprendan de forma automática con la finalidad de resolver las ambigüedades. Los métodos débilmente supervisados, de alguna manera siempre necesitan de algún tipo de aprendizaje aunque en menor escala para realizar su labor. Por otra parte, los métodos no supervisados no necesitan ningún texto etiquetado manualmente. Por último, se han explicado los métodos basados en conocimiento, los cuales se sirven de recursos externos para acometer su tarea. Es importante mencionar que muchos de estos sistemas se mezclan y a veces es difícil identificar a qué grupo pertenecen. Se reconocen entonces a nivel general dos tipos de sistemas: supervisados (aquí se incluyen los que aplican aprendizaje) y no supervisados (aquí se incluyen el resto).

Para evaluar la confianza que ofrecen los sistemas de WSD en otras posibles aplicaciones finales, se han citado las competiciones de Senseval. Con el fin de obtener mayores detalles de lo que en estas competiciones ha ocurrido, se ha hecho un análisis exhaustivo de los resultados obtenidos por los sistemas participantes y sumado a ello, se han añadido las descripciones de los mejores sistemas.

Tras el análisis de las distintas competiciones, ha quedado patente que los sistemas supervisados siempre se sitúan en las primeras posiciones. Pero a medida que se desarrollan estos eventos se hace más pequeña la diferencia entre las puntuaciones de los métodos supervisados y los que no lo son. Esto es debido a que se han ido introduciendo aproximaciones basadas en conocimiento, que al explorar las interconexiones semánticas de los recursos, logran establecer con altos grados de fiabilidad sus respuestas. En los comienzos, los sistemas utilizaban estructuras de árboles para obtener conocimiento y en la actualidad los sistemas pueden utilizar complicadas estructuras de grafos semánticos. Se puede demostrar que al reducir la granularidad de los inventarios de sentidos, las puntuaciones se elevan considerablemente. Esto no indica que esté resuelto el problema de la desambiguación de los sentidos de las palabras. Otra de las observaciones importantes tras haber analizado el comportamiento de los sistemas en las competiciones, es que, el *baseline* MFS siempre ocupa los primeros puestos del *ranking*. Este no contiene ninguna información semántica pero incluso así sus respuestas son relativamente buenas. En estas competiciones el uso de los recursos semánticos se ha manifestado de la siguiente manera:

- **WordNet**, como base de conocimiento lo utilizan sistemas como Sinequa-LIA-HMM, USM, ITT, DIMAP en Senseval-2. Propuestas como (Izquierdo *et al.*, 2007) (Izquierdo *et al.*, 2010) (Montoyo Guijarro, 2002) también se sirven del mismo recurso. Aunque Izquierdo genera su propio recurso de Clases Semánticas.
- **WND**, varias propuestas como (Vázquez et al., 2004b, Strapparava et al., 2004, Magnini et al., 2008, Magnini and Cavaglia, 2000, Buscaldi and Rosso, 2007, Ion and Stefanescu, 2010), utilizan solamente los dominios de (Magnini *et al.*, 2008) mediante la detección de Dominios Relevantes en textos.
- **SUMO**, lo usan aproximaciones en (Zouaq et al., 2009, Villarejo et al., 2005).
- ***eXtended* WordNet**, en (Naskar and Bandyopadhyay, 2007, Martinez et al., 2007, Agirre and Soroa, 2009, Sinha and Mihalcea, 2007).
- **FrameNet**, es aplicado por (Laparra *et al.*, 2010) y otros.
- **MCR**, es aplicado por el sistema *Meaning* (Villarejo *et al.*, 2004) y (Martinez et al., 2007).
- **eXtendedWordNet** y **MRC**, lo combina (Martinez et al., 2007).

- La combinación de **WND**, **SemCor** y **LDC-DCO**[63] es aplicado por (Navigli and Velardi, 2004), se debe destacar que es válida la combinación pero SemCor y LDC-DCO son corpus.
- **WordNet**, **WND**, **SUMO** y **WordNet Affects**, son combinados por *Relevant Semantic Trees* y *Relevant Cliques* por UMCC-DLSI en Semeval-2 (Gutiérrez *et al.*, 2010b) . Esta propuesta de integración forma parte del contenido de esta Tesis, figurando como una primera propuesta de integración de recursos (en detalle véase el Capítulo 4).

Finalmente, se puede extraer como conclusión que muchos sistemas basados en conocimiento obtienen altos resultados al aplicar el análisis desde el punto de vista de una o dos dimensiones semánticas. También, que el uso de la frecuencia de los sentidos obtiene siempre resultados relevantes. Por esta razón, se reitera la propuesta principal de la Tesis orientada a desarrollar recursos y métodos basados en conocimiento soportados sobre la base de la Semántica Multidimensional, capaces en conjunto, de superar los resultados alcanzados por sistemas sin supervisión de WSD y añadiendo además la posibilidad de aplicarlos en el área de la Minería de Opiniones. En los siguientes capítulos, primero se describe un nuevo recurso semántico multidimensional, que sirve de base para varias aproximaciones de WSD y en los siguientes la aplicación de la semántica multidimensional en otras tareas finales del PLN, a fin de resolver tipos diferentes de ambigüedades del lenguaje (Ratnaparkhi, 1998).

[63] http://www.ldc.upenn.edu/

3. INTEGRACIÓN DE RECURSOS SEMÁNTICOS

En este capítulo se expone el desarrollo de un recurso semántico que incorpora diferentes bases de conocimiento. Entre las fuentes integradas se encuentran WN, WND, WNA, SUMO, SC y SWN. Esto posibilita visualizar el entorno de las palabras y sus sentidos desde perspectivas muy diferentes. Para su creación se plantean dos fases, en la primera se integran WN, WND, WNA, SUMO y en la segunda SC y SWN. Finalmente se documenta el grado de fiabilidad de la herramienta, abordando ciertas facilidades de las que algunas aproximaciones del PLN han hecho uso.

3.1. INTRODUCCIÓN

En la mayoría de tareas de Procesamiento del Lenguaje Natural (PLN) es necesaria la utilización de recursos externos tales como: diccionarios, tesauros, ontologías, etcétera. Estos recursos proporcionan sus respectivas estructuras internas, interfaces, relaciones entre conceptos y otras características. Entre todos los recursos desarrollados, uno de los más utilizados en sus diferentes versiones e idiomas es WN. Debido a su gran repercusión, otros recursos tales como WND (Magnini and Cavaglia, 2000), SUMO (Niles and Pease, 2001), WNA (Esuli and Sebastiani, 2006) y otros, han sido desarrollados basados en las relaciones y estructuras internas de WordNet.

Actualmente, el desarrollo de tareas para la clasificación de documentos, discriminación de entidades, detección de autoría, entre otros, ha hecho patente la necesidad de disponer de ciertos recursos semánticos que proporcionen información adicional a los contextos analizados (ej. detección de subjetividad, dominio contextual, etc.). El principal problema en el uso de estos recursos es su descentralización. A pesar de que la mayoría se basa en las relaciones internas de WN, no comparten una interfaz común que pueda proporcionar información de forma cohesionada.

Como ya se había mencionado en la sección 2.4.3, algunos autores han unido fuerzas para construir redes semánticas con una interfaz común, como MWN (Pianta *et al.*, 2002), EWN (Dorr and Castellón, 1997), MCR (Atserias *et al.*, 2004) y otros. La deficiencia de ellos con respecto al objetivo de este capítulo es que en su mayoría la integración fundamental es de carácter léxico (aplican alineamiento de idiomas), aunque también incorporan unos pocos recursos conceptuales. Se plantea la necesidad de conocer: ¿Qué recursos semánticos se pudieran integrar para poder aplicar Análisis Semántico Multidimensional en tareas del Procesamiento del Lenguaje Natural?

Luego de analizar diferentes recursos léxicos y conceptuales, e identificar que WN sirve como interfaz de integración, se plantea como objetivo desarrollar una herramienta que permita enlazar diferentes recursos basados en WN y explotar todas sus relaciones: hiperonimia, meronimia, sinonimia, etc. Para ello, se toma como enfoque la integración paulatina (en cada versión de la herramienta) de todos los recursos semánticos que se alinean a este. Requiriendo que la herramienta resultante proporcione una interfaz común, a través de una red de nodos interconectados entre sí.

3.2. PRIMERA FASE DE INTEGRACIÓN

El proceso de integración se ha centrado en esta fase, únicamente para WN, WND, SUMO y WNA. Pretendiendo la obtención de una herramienta final que aporte ayuda y fiabilidad para cuando se realice una única consulta a este conjunto. A continuación se describen los elementos que se han tenidos en cuenta para la integración semántica.

Es necesario comentar que WN es uno de los recursos más utilizados en PLN, debido a ello, diversos autores han desarrollado taxonomías que añaden nueva información y proporcionan nuevas relaciones entre conceptos (Magnini and Cavaglia, 2000, Valitutti et al., 2004, Niles and Pease, 2003, Niles, 2001, Sara and Daniele, 2009, Forner, 2005, Strapparava and Valitutti, 2004)). Entre los recursos derivados de WN cabe destacar SUMO (Niles and Pease, 2001) , WND (Sara and Daniele, 2009) y WNA (Strapparava and Valitutti, 2004). Tras la aparición de estos, varios autores (Gliozzo et al., 2004, Magnini et al., 2002, Magnini et al., July 2002, Vázquez, 2009, Vázquez et al., 2004a) entre otros, han desarrollado nuevos métodos y sistemas que utilizan este enriquecimiento. Estos han demostrado mejoras en tareas como: extracción de información, elaboración de resúmenes, indexado de documentos o desambiguación léxica. Es de destacar que cada uno de estos autores basan sus aproximaciones en el uso de un recurso o máximo dos (véase la sección 2.7). Esto es debido a que no existe ninguna herramienta que integre todos los recursos mapeados con WN, y que facilite el trabajo de los investigadores. Luego de analizar dicha necesidad, se ha considerado la posibilidad de integrar la mayor cantidad de recursos léxicos y ontológicos en una única herramienta de utilidad para toda la comunidad científica.

En esta primera versión se propone utilizar como núcleo común WN, ya que, su estructura interna y relaciones entre conceptos proporciona información relevante a diversas tareas de PLN. En la Figura 11 se ilustra el esquema lógico al cual se le conoce como "*Integration of Semantic Resources based in WordNet"* (ISR-WN) (Gutiérrez *et al.*, 2010a). En ella se puede apreciar cómo varias dimensiones conceptuales se entrelazan a partir de un núcleo común.

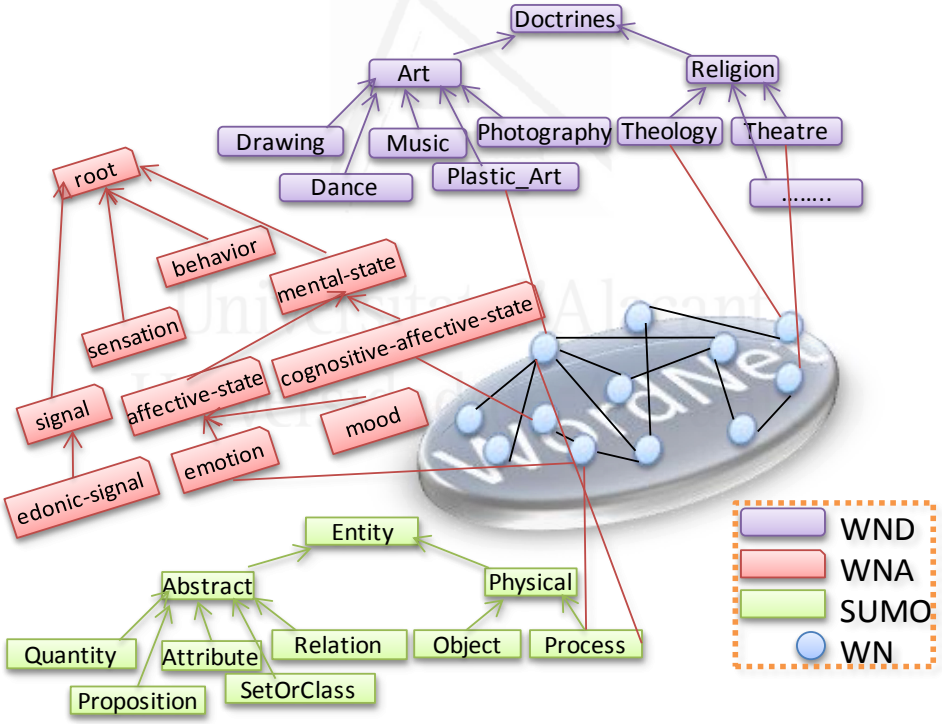


Figura 11. Integración de recursos semánticos basados en WordNet (ISR-WN).

3.2.1. PROPUESTA DE INTEGRACIÓN

En este apartado se describen las características del modelo utilizado para la integración de recursos. Como se ha mencionado anteriormente, el esquema de integración, toma como núcleo a WN y vinculado a este los recursos SUMO, WND y WNA. Cada uno se presenta con sus peculiaridades e incluso teniendo en cuenta algunas de sus versiones. Como todos los recursos utilizados están anotados en idioma inglés, solamente se realiza la integración para este idioma.

Partiendo del esquema de la Figura 12 se obtiene como resultado un recurso capaz de integrar las dimensiones semánticas implicadas, que anteriormente se encontraban accesibles de forma individual. En la Figura 12 se muestra cómo los *synsets* se representan por palabras y a su vez se relacionan con las distintas jerarquías de SUMO, WND y WNA a través de diferentes ficheros de mapeos. Estos permiten relacionar distintas versiones en las que están anotados los recursos, obteniendo como resultado un grafo semántico, muy útil en aplicaciones de PLN.

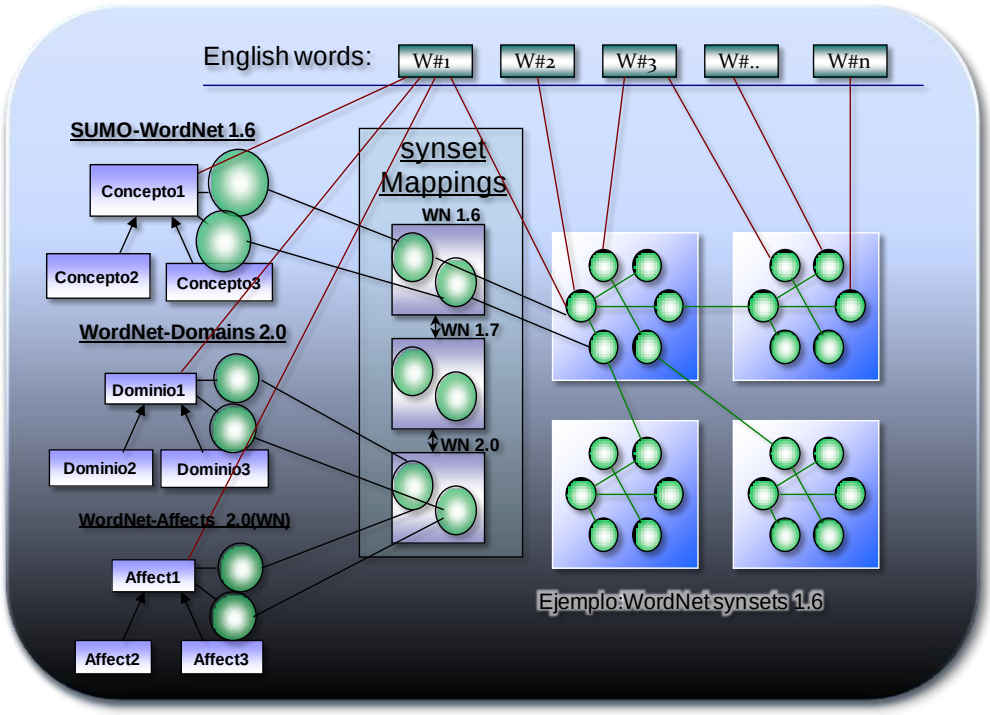


Figura 12. Arquitectura de ficheros para la integración semántica de WN.

Como se puede observar las palabras están representadas tanto en los sentidos de WN como en los conceptos que conforman las taxonomías de SUMO, WN y WNA. Para permitir la integración y creación de una gran red de conocimiento de varias dimensiones conceptuales, se utilizan, según la versión en que estén anotados los recursos, mapeos de diferentes versiones de WN. En algunos casos, existen relaciones que no están contempladas en todas las versiones. Para solucionar este problema se ha propuesto navegar a través de tantas versiones como sea necesario hasta conseguir todos los datos necesarios. Es importante destacar que en esta fase se propone el uso de WN 1.6 como núcleo y no otras versiones de WN.

3.2.2. *SOFTWARE* Y LIBRERÍAS DE CLASES

Una vez realizado el estudio de las relaciones y sus conexiones entre los diferentes recursos, se propone la creación de un software específico que incorpora librerías de clases de programación, capaces de navegar en el interior del grafo semántico creado. Los nodos presentes en la red se implementan utilizando las jerarquías de clases que se observan en el diagrama de la Figura 13, con el fin de que todos los recursos coexistan por igual.

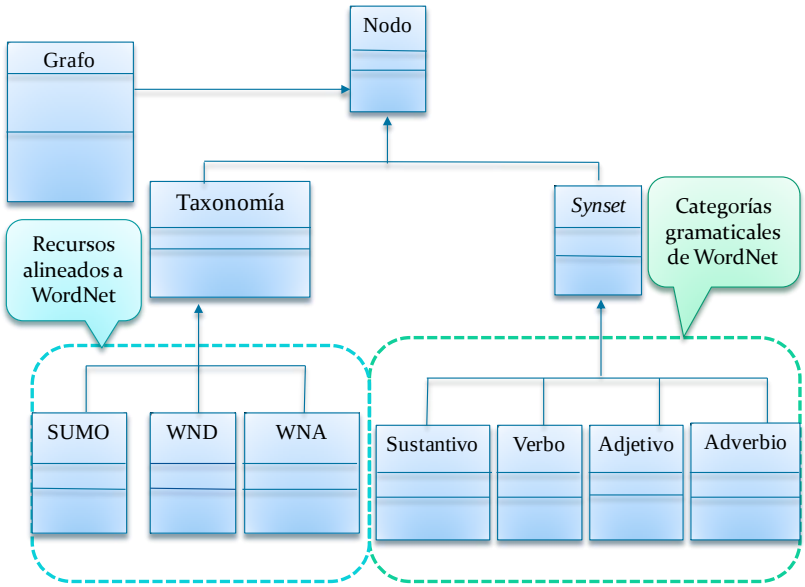


Figura 13. Jerarquía de clases de las librerías del recurso Integrador.

Con esta representación del diagrama de la Figura 13, es posible conceptualizar los diferentes orígenes de datos (SUMO, WND, WNA, sustantivo, adjetivo, verbo y adverbio); estos cuatro últimos representan los *synsets* correspondientes a las categorías gramaticales. Todos los nodos, de forma general contienen diferente información pero la navegación se realiza a través de todos ellos. Como resultado de la implementación de las clases se genera una API (*Application Program Interface*) capaz de recuperar los nodos identificados en la red según la palabra de búsqueda, al utilizar técnicas estructuras de datos tales como caminos mínimos, recorridos en profundidad y anchura, descritos en (Cormen et al., 1990, Díaz, Aho et al., 1974).

El esquema planteado en la Figura 12 toma como núcleo la versión 1.6 de WN. Esta decisión no implica que en el futuro no se use otra. Como resultado, el modelo integrador respeta las relaciones existentes entre los *synsets* de WN. Además, entre los conceptos de SUMO[64] se conservan las relaciones que proveen los mapeos con WN. Las relaciones que se establecen entre etiquetas propias de WND y WNA son las de hiponimia e hiperonimia, pero la relación de los *synsets* con estas es de pertenencia.

Las conexiones de la red de conocimiento obtenidas, permiten navegar a través de todos los enlaces de los recursos integrados. Por ejemplo, para el *synset* 00124616 de la palabra *run* por la relación de hiponimia se obtiene:

- 00124616 *run*
  - hiponimia (00124895 *earned run*)
  - hiponimia (00125039 *unearned run*)
  - hiponimia (00125179 *run batted in*)
  - hiponimia (*maneuver* [SUMO])

De no seleccionarse un tipo específico de relación se navegaría por todas ellas.

[64] http://suo.ieee.org/SUO/SUMO/index.html

3.3. SEGUNDA FASE DE INTEGRACIÓN

En esta fase se propone además de incrementar dos dimensiones más a la integración presentada: las SC de (Izquierdo *et al.*, 2007) y SWN (Esuli and Sebastiani, 2006), implicar a WN1.6 y WN2.0. La Figura 14 presenta el modelo lógico por el cual se dirige esta nueva propuesta. Tal y como se observa cada dimensión (recurso) está conectada a WN (el núcleo de la red) a través de sus interconexiones. Según se ha explicado en la sección 2.4.2, existen diferentes versiones de WN y también diferentes versiones de los demás recursos mapeados a versiones de WN. Se hace es necesario, adecuar los mapeos de cada versión con cada recurso. Esta propuesta incluye ahora seis recursos: WND, WNA, SUMO, SWN, SC y WN. La cual se le conoce como "*Enriching the Integration of Semantic Resources based on WordNet*" (*Enriched* ISR-WN) (*Gutiérrez et al., 2011a*).

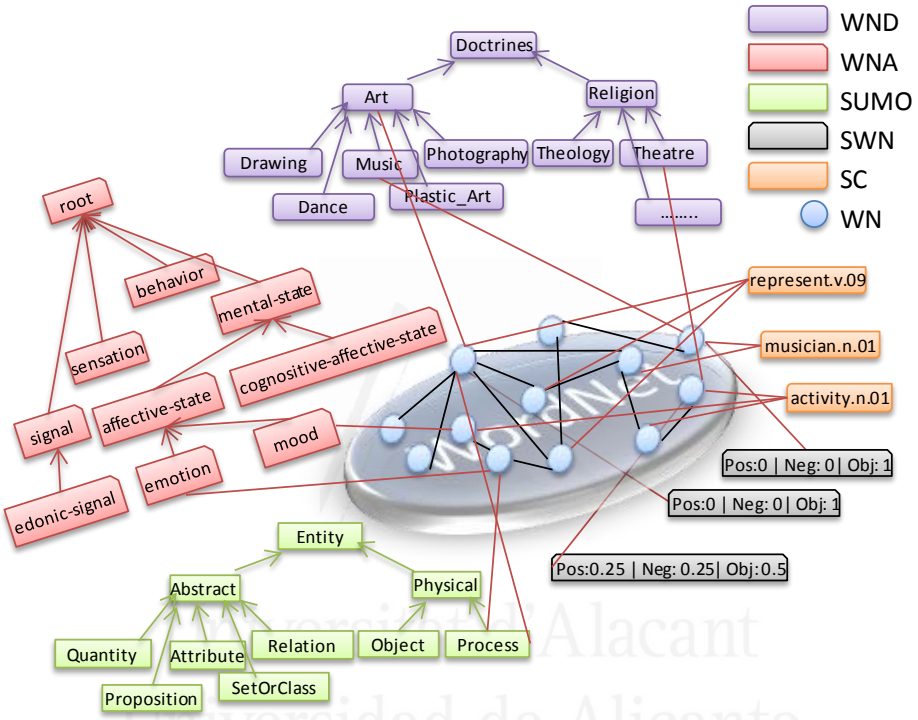


Figura 14. Modelo lógico del enriquecimiento de la integración de recursos (*Enriched* ISR-WN).

3.3.1. PROPUESTA DE INTEGRACIÓN

Esta sección describen las características del modelo usado para obtener la integración de diferentes recursos. El esquema de integración toma a WN como núcleo y lo vincula con cada uno. Como nota aclaratoria, es necesario conocer que en dependencia del núcleo deseado (WN 1.6 0 WN 2.0) para despertar la herramienta, se genera la red de conocimiento con particulares características. Para ello, se han tenido en cuenta cada una de las peculiaridades de cada dimensión implicada. Debido a que todos los recursos semánticos se encuentran etiquetados en idioma inglés, se reitera que la integración únicamente tiene en consideración este idioma.

Del modelo presentado en la Figura 14, una nueva arquitectura se ha obtenido. Esta incluye la integración de los recursos de la primera versión, además de la posibilidad de acceso a cada uno de ellos de modo individual con la adición de dos dimensiones más. La Figura 15 muestra como los *synsets* están representados por palabras y al mismo tiempo sus relaciones con varias taxonomías (SUMO, WND y WNA) además de SC y SWN. Los vínculos entre todos con WN se realizan a través de diferentes ficheros de mapeos. Estas relaciones permiten enlazar diferentes versiones en las han sido anotados los recursos, obteniendo como resultado, un

manipulable grafo semántico válido para ser utilizado por aplicaciones del PLN. Como se puede apreciar en la Figura 15, todas las etiquetas de los recursos mencionados ((taxonomías de SUMO (color verde), WND (color malva) y WNA (color rojo); etiquetas de SC (color naranja) y descripciones de SWN (color gris)) se asocian a *synsets* de WN.

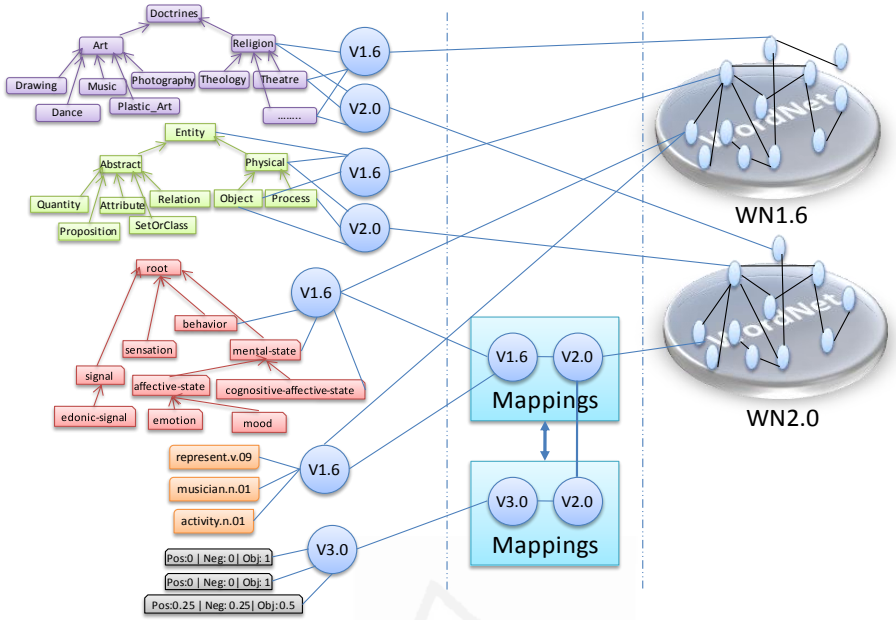


Figura 15. Arquitectura de ficheros para el enriquecimiento de integración semántica de WN.

Con respecto a integrar una gran red de conocimiento proveniente de varias dimensiones conceptuales, se utilizan mapeos de WN de diferentes versiones. En algunos casos existen relaciones que no están contempladas en todas las versiones. Para resolver ese problema, se propone la navegación a través de ficheros de versiones de WN con el fin de obtener los datos necesarios. Por ejemplo, WNA está vinculado con WN 2.0, SWN está vinculado con ambas versiones (WN1.6 y 2.0), y las SC está vinculado con WN 2.0. Esta es la naturaleza de los recursos, ahora entonces la propuesta tiene que navegar a través de versiones de WN y lograr la integración de todas. Se debe destacar que en esta versión del recurso de integración se tienen en consideración para WND y SUMO versiones alineadas con WN 1.6 y WN 2.0. Lo cual posibilita reducir el riesgo de inexistencia de *synsets* en los ficheros de mapeo.

El modelo presentado en la Figura 15 ofrece la posibilidad de elegir como núcleo WN 1.6 o 2.0 dependiendo del objetivo del usuario. Esta decisión no limita que otras versiones de WN sean implementadas en el futuro. Como se puede observar, en comparación con el modelo de la primera fase y el de esta, no solamente se incorporan nuevos recursos, sino que también se puede elegir la versión de WN con la cual trabajar.

Como resultado, el modelo de integración respeta el nombre de las relaciones existentes entre los *synsets* de WN *Princeton.* Las categorías de SUMO, también conservan las relaciones con los mapeos con WN. Sin embargo, las relaciones establecidas entre las taxonomías de WND y WNA se manifiesten ahora como de hiponimia e hiperonimia, pero entre estos recursos y los *synsets*, el nombre de asociación es de membrecía (ej. un *synset* puede pertenecer a una/varias etiqueta/s de WND/WNA).

Es importante subrayar que la versión de WNA 1.1 está poblada con nuevas relaciones entre *synsets* que no existen en WN *Princeton* ni en la versión WNA 1.0 (ej. *entailment*, *cause*), estos nuevos lazos permiten vincular verbos, adjetivos y adverbios a los sustantivos. Estas consideraciones son tomadas en cuenta para el desarrollo de la nueva propuesta de integración en contraste con la primera fase.

Las conexiones de la red de conocimiento obtenidas permiten navegar entre todas las relaciones de los recursos integrados. Por ejemplo, tomando en consideración la palabra del inglés *atrocious* y utilizando la versión 1.6 de WN, se obtienen una lista de *synsets* pertenecientes *atrocious*. Como ejemplo se recupera la siguiente información para uno de sus sentidos *atrocious#3*:

- ***Sense**: alarming#1 **Relation**: Similar-To*
- WND*: Psychological_Feature **Relation**: Pertainym*
- SUMO*: SubjetiveAssessmentAttribute **Relation**: Hyponym*
- WNA*: Emotion **Relation**: Pertainym*
- WNA*: Horror **Relation**: Pertainym*
- **S**WN*: atrocious#3 **Pos**: 0|**Neg**: 0.625 |**Obj**: 0.375 **Relation**: SentiWN-Description*
- ***Sense:** horror#1 **Relation**: Cause*

Además de obtener la información mostrada, se consigue para cada sentido la numeración del *offset*, la categoría gramatical, la lista de sinónimos y las glosas, sumándole a esto las descripciones particulares de cada recurso con el que se asocie el *synset* . También, el buscador de la herramienta integradora plantea que si la palabra de entrada coincide con alguna etiqueta de algún recurso, esta, igualmente es obtenida. Sin embargo, para cada etiqueta obtenida se pueden recuperar todas las etiquetas de cualquier recurso que se encuentren relacionadas. Nótese, que originalmente la etiqueta de WNA *Horror* no está vinculada directamente con *atrocious#3*, pero en la propuesta (*Enriched* ISR-WN) se asume que si un *synset* (ej. *atrocious#3*) se vincula con un sustantivo (ej. *horror#1*) mediante una relación afectiva obtenida de WNA1.1, este sentido también será vinculado con las etiquetas de WNA con el que el sustantivo se asocie. En este caso el *synset horror#1*, se asocia directamente a la etiqueta de WNA *Horror*, entonces *atrocious#3* también se vinculará asumiendo la relación del *synset* vinculado directamente.

3.3.2. LIBRERÍAS DE CLASES

Debido a que esta nueva propuesta integra dos recursos más que no concebidos en el diseño inicial, entonces, se le agregan dos clases al diseño (SWN para las descripciones de sentimientos y SC para las etiquetas de clases semánticas). En la Figura 16 se muestra el nuevo diseño de clases. En la figura no se perciben detalles de métodos, propiedades, etc., solamente para obtener una idea general. Este diseño posibilita aplicar un conjunto de funciones definidas en la clase grafo a todos los recursos por igual.

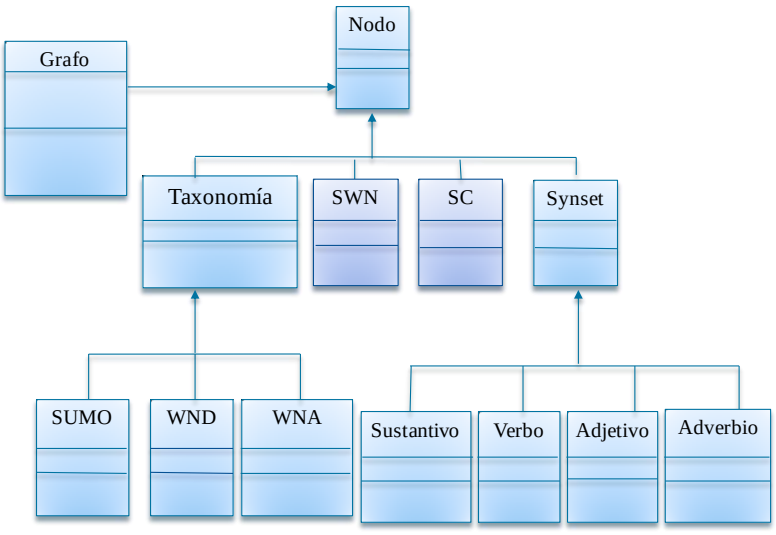


Figura 16. Diseño de clases del recurso ISR-WN enriquecido.

3.4. ANÁLISIS DE RESULTADOS

En esta sección se analizan los diferentes resultados obtenidos a partir de haber desarrollado esta gran base de conocimiento.

3.4.1. RESULTADOS DE LA PRIMERA FASE DE INTEGRACIÓN

Con el objetivo de valorar la fiabilidad de la primera versión, se realizan dos etapas de evaluación. Una primera etapa para evaluar la extracción de conceptos de WN. En esta etapa, se recuperan diferentes *synsets* junto con todas sus relaciones: hiponimia, hiperonimia, etc., y se comparan los resultados obtenidos con las herramientas de navegación de WordNet originales para la versión 1.6.

Luego de haber evaluado la fiabilidad de recuperación de *synsets* con un 100% de coincidencias respecto a sentidos y relaciones recuperadas, se procede con una segunda etapa. Esta se realiza con el objetivo de determinar el grado de fiabilidad del módulo con respecto al resto de recursos integrados: WND, WNA y SUMO. En esta segunda etapa se desarrollan tres pruebas:

1. Medición manual de precisión entre las recuperaciones de conceptos de SUMO del módulo y las recuperadas en los ficheros de mapeos proporcionados por el sitio oficial de SUMO[65].
2. Medición manual de precisión entre las recuperaciones de conceptos de WND del módulo y las recuperadas en los ficheros de mapeos proporcionados por (Sara and Daniele, 2009)[66].
3. Medición manual de precisión entre las recuperaciones de conceptos de WNA del módulo y las recuperadas en los ficheros de mapeos proporcionados por los mismos autores de WND.

Para la extracción de las muestras tomadas en el desarrollo de todas las pruebas, se utiliza la fórmula estadística denominada "Muestra Representativa" (Fernández, 1996) descrita en la ecuación (30). Donde N es la población, K es intervalo de confianza, E es el error, P es la probabilidad de éxito y Q es la probabilidad de fracaso.

$$M = \frac{N * K^2 * P * Q}{E^2(N-1) + K^2 * P * Q} \tag{30}$$

Para los tres experimentos se toman los siguientes valores:

$$K = 0.95, E = 0.05, P = 0.5 \text{ y } Q = 0.5$$

Únicamente varía el tamaño de la población para cada caso, dependiendo de la cantidad de conceptos que posee cada dimensión evaluada.

El primer análisis que respecta a SUMO, $N = 568$ conceptos, recopilando una muestra de 78 conceptos. Los resultados muestran un 100% de fiabilidad. En este caso, no existe margen de error debido a los conceptos de SUMO (anotado para WN1.6) y sentidos de WN 1.6 tienen una relación $1 - 1$.

[65] http://suo.ieee.org/SUO/SUMO/index.html
[66] http://wndomains.fbk.eu/download.html

En el segundo análisis implica a WND, con $N = 170$ conceptos, se calcula una muestra de 59 conceptos, de ellos se recuperan una cantidad de 56. En este caso la extracción se realiza correctamente al 94.9%. Luego de un análisis profundo de esta inconsistencia, se detectan fallos en las traducciones entre versiones de WN. Esto significa que en los mapeos de un *offset* de la versión 2.0 con uno de la versión 1.6 existe un margen de error de un 5%.

En el tercer análisis se enfoca en WNA con $N = 304$ conceptos, se calcula una muestra a comparar de 70 conceptos. De ellos se recuperan con similitud en las respuestas una cantidad de 67 conceptos, obteniendo un 95,7% de precisión. El 4% de error aproximado es debido a inconsistencias en los ficheros de mapeos entre versiones y los *offsets* reales de los ficheros de sentidos.

Como resultado de estos mapeos se obtiene un software capaz de recuperar de una palabra de entrada, los conceptos de SUMO, WND, WNA y WN del módulo integrador. En Figura 17 se visualiza un ejemplo de navegación con la palabra de entrada *mouse*.

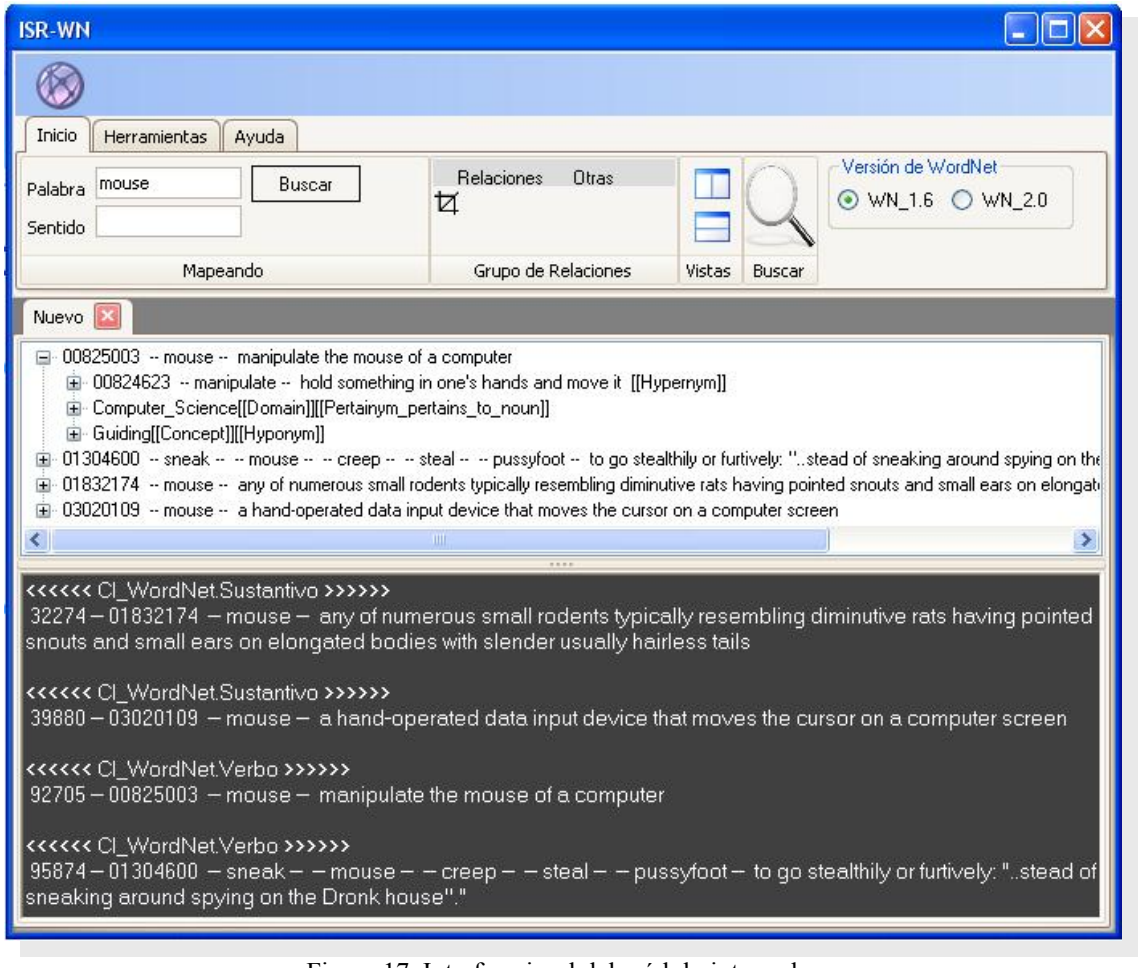


Figura 17. Interfaz visual del módulo integrador

Esta interfaz es capaz de detectar multipalabras para mejorar la búsqueda en la red de nodos. Por ejemplo, si se desea conocer qué nodos de la red se asocian con el dominio *Computer_Science* recuperado, basta con hacer *clic* sobre él y se desplegará todo un árbol asociado a ese dominio con sus respectivas relaciones. Además, se puede ver cómo los sentidos recuperados son categorizados según WN como nombres (sustantivos), adjetivos, verbos y adverbios, junto con los conceptos de SUMO, WND y WNA. En la imagen se tiene seleccionada la versión de WN 1.6 debido a ser la versión lista para esta etapa, estando preparada la interfaz para el uso también de WN 2.0, introducido en la segunda fase de desarrollo del recurso.

Este software se proporciona una API que tiene un conjunto de funcionalidades muy útiles en diversas tareas de PLN. Por ejemplo, los autores (Banerjee and Pedersen, 2002, Montoyo Guijarro, 2002) tratan el problema de la desambiguación léxica utilizando solamente WN, otros autores van un poco más allá e introducen los dominios de WND en las tareas de desambiguación del sentido de las palabras (WSD) como son: (Magnini et al., 2002, Vázquez, 2009, Magnini et al., July 2002). Un estudio realizado por (Gutiérrez, 2010) demuestra que la incorporación de los conceptos de SUMO, WND y WNA a un sistema de desambiguación mejoran los resultados obtenidos en la tarea de Senseval-2. Cabe destacar que entre los sistemas participantes superados muchos utilizaban WN y WND, pero ninguno tenía en cuenta los cuatro implicados en esta versión de la herramienta.

3.4.2. RESULTADOS DE LA SEGUNDA FASE DE INTEGRACIÓN

Luego del desarrollo del enriquecimiento del recurso, se conduce a la elaboración de análisis de las cantidades de *synsets* que deberían ser alineados con respecto a los enlaces que realmente se efectúan. En este análisis no se obtiene una muestra representativa como en la fase anterior, sino que se analiza la totalidad de los elementos implicados. En la Tabla 14 se muestran las cantidades de etiquetas implicadas al utilizar el recurso con WN 1.6 como núcleo, además de la totalidad de *synsets* a ser alineados y los que lo logran, seguidos de un cálculo de porcentaje. Se debe anotar que en la primera fase algunas etiquetas de SUMO y WNA se extraviaron, debido a que el sistema no era capaz de reconocerlas. Ahora en esta versión se intenta asumir la gran mayoría y las que no se incorporan se justifican más adelante.

Como se puede apreciar la reducción del uso de ficheros de alineamiento provoca que WND, SUMO y SC se alineen completamente. Esto es posible porque los tres recursos usados fueron construidos sobre WN 1.6.

	WND2.0	**SUMO**	**WNA_1.0-1.1**	**SC**	**SWN_30**
# Etiquetas	170	569	309	1231	117659
Synsets a alinear					
n	86901	67923	1256	66025	82114
a	19322	18531	2418	-	18157
v	12843	12469	801	12127	13767
r	3735	3627	614	-	3621
Total de synsets a alinear	122801	102550	5089	78152	117659
Synsets alineados					
n	86901	67923	1096	66025	56563
a	19322	18531	2125		8757
v	12901	12469	474	12127	9223
r	3735	3627	549		2101
Total de synsets alineados	122801	102550	4244	78152	76644
Diferencia	0	0	845	0	41015
% alineado	**100.00**	**100.00**	**83.40**	**100.00**	**65.14**

Tabla 14. *Synsets* alineados a cada recurso utilizando WN1.6 como núcleo.

La diferencia de 845 en la Tabla 14 entre WNA 1.0 y WNA 1.1 (respecto a los *synsets* alineados y los que deberían alienarse) es debido a que se han utilizado las etiquetas emocionales de WNA 1.1. En este caso en especial, la mayoría de las etiquetas en WNA 1.0 se mantienen en la versión 1.1 de WNA, pero existen algunas de ellas que no se hallan en WNA 1.0. Entonces los *synsets* que deberían ser alineados con estas etiquetas que no persisten en la versión 1.1, pierden sus vínculos dentro de la integración de recursos. Por ejemplo, las etiquetas que se muestran a continuación no están presentes en la taxonomía de WNA en *Enriched* ISR-WN*: "attitude", "emotional response", "psy", "man", "sympathy", "sta", "softheartedness", "joy-pride", "identification", "levity-gaiety", "general-gaiety", "empathy", "positive-concern", "compatibility", "kindheartedness" and "buck-fever"*. Por esta razón, el recurso de integración no incluye los vínculos entre estas etiquetas y los *synsets*. Es importante resaltar que las

taxonomías utilizadas en el recurso de integración son las más recientes de WND y WNA (ej. WND 3.2, WNA 1.1) y que algunas etiquetas no pudieron ser incluidas por motivos de codificación.

Se tienen en cuenta para el alineamiento de WNA los ficheros de mapeos de ambas versiones, con el objetivo de asociar la mayor cantidad de etiquetas emocionales a los *synsets*. WNA 1.1 presenta una peculiaridad, que solamente los sustantivos están enlazados a etiquetas emocionales. Pero se debe anotar que WNA 1.1 ha propuesto la generación de nuevas relaciones entre *synsets* de WN (ej. *entailment*, *cause*), indicando que se pueden obtener nuevos vínculos afectivos entre verbos, adjetivos y adverbios con los sustantivos. En esta fase de integración semántica ha decidido tener en cuenta también estas relaciones dentro del ISR-WN.

La principal diferencia entre los alineamientos pertenecientes a SWN, es que este recurso ha sido desarrollado sobre WN 3.0. Entonces muchos sentidos que existen en WN 3.0 no lo están en WN 1.6. Esto justifica la pérdida de vínculos mostrados en la Tabla 14. La Tabla 15 muestra la misma comparativa que la Tabla 14 pero ahora tomando como núcleo WN 2.0.

	WND 3.2	**SUMO**	**WNA_1.0-1.1**	**SC**	**SWN_30**
# Etiquetas	170	569	309	1231	117659
Synsets a alinear					
n	103504	79688	1256	66025	82114
a	19398	18564	2418	-	18157
v	19398	13507	801	12127	13767
r	3835	3663	614	-	3621
Total de synsets a alinear	146135	115422	5089	78152	117659
Synsets alineados					
n	103504	79688	1089	65904	78061
a	19398	18564	2118		11052
v	19398	13507	473	12064	13207
r	3835	3663	580		3428
Total de synsets alineados	146135	115422	4260	77968	105748
Diferencia	0	0	829	184	11911
% alineado	**100.00**	**100.00**	**83.71**	**99.76**	**89.88**

Tabla 15. *Synsets* alineados a cada recurso utilizando WN 2.0 como núcleo.

Como se observa SC ahora ha perdido varios vínculos, esto es debido a que este recurso ha sido desarrollado sobre WN 1.6 y al asociarse con WN 2.0 es necesario utilizar los ficheros de mapeos donde siempre existe la posibilidad que sentidos que existen en 2.0 no lo estén en 1.6. Por otra parte, se puede apreciar que los vínculos con SWN se incrementan mucho más, debido a que al utilizar los mapeos entre una sola versión se reducen las inexistencias de *synsets*.

En ambas tablas no se juzgan a WNA 1.0 y WNA 1.1 por separado, en esta propuesta se fusionan las dos versiones. Todas las etiquetas de WNA 1.0 que se mantienen en la taxonomía de WNA 1.1 se encuentran alineadas a todos los *synsets* que WNA 1.0 propone, además de los vínculos que el propio WNA 1.1 sugiere. Al utilizar esta vinculación especial, el recurso de integración semántica es capaz de facilitar la coexistencia de ambas versiones.

3.5. INVESTIGACIONES QUE HAN UTILIZADO ISR-WN

Entre los principales propósitos del desarrollo del recurso de integración semántica, está ayudar en diferentes tareas del Procesamiento de Lenguaje Natural desde una perspectiva multidimensional. Los atributos semánticos que ofrece este recurso, han posibilitado que variadas investigaciones hagan uso de él. A continuación se citan algunas de las contribuciones en las que ha sido partícipe:

- En la participación del grupo de investigación UMCC-DLSI en la competición de Semeval-2010 para la tarea 17: *All-words Word Sense Disambiguation on Specific Domain* (Agirre et al., 2010). Donde se identifican Árboles Semánticos Relevantes

(*Relevant Semantic Trees* (RST)) de las frases para resolver problemas de ambigüedad semántica (Gutiérrez et al., 2010b).

- En la conferencia del RANLP 2011[67] combinando ISR-WN (WN, WND, WNA y SUMO) y con la frecuencia de sentidos, esta propuesta supera los resultados de la anterior y muchas otras de escala mundial (Gutiérrez *et al.*, 2011b).
- En la conferencia NLDB2011[68] con una aproximación basada en grafos para la tarea de WSD, teniendo como base de procesamiento el recurso de integración semántica (Gutiérrez et al., 2011d). Esta propuesta alcanza lugares prometedores en evaluaciones realizadas sobre el sistema de competición Senseval-2 (Cotton *et al.*, 2001).
- Otras temáticas en las que se ha aplicado la visión multidimensional del recurso, es en la Minería de Opiniones (*Opnion Mining*). Por ejemplo el trabajo presentado en WASSA'11[69] utiliza el recurso para evaluar tres tareas del análisis de opiniones del sistema de competiciones MOAT (*NTCIR Multilingual Opinion Analysis Task*[70]) respecto a medir la clasificación de frases de acuerdo la existencia de opinión, si es relevante a una pregunta del tópico, y la polaridad de la opinión. La principal idea de esta propuesta, busca extraer conceptos relevantes como características distintivas de las frases, y asociarles puntuaciones de polaridades presentes en SWN de *Enriched* ISR-WN. Esta obtiene relevantes resultados que la pudieran colocar entre los primeros puestos de la competición MOAT (Gutiérrez *et al.*, 2011c).
- En la actualidad el Departamento de Lenguajes y Sistemas Informáticos (DLSI) de la universidad de Alicante (España) y el Departamento de Informática de la Universidad de Matanzas (Cuba), enfocan en multidisciplinares investigaciones con el uso de este recurso.

3.6. CONCLUSIONES

Como resultado de este capítulo se ha desarrollado una herramienta capaz de integrar recursos semánticos para poder aplicar Análisis Semántico Multidimensional en tareas del Procesamiento del Lenguaje Natural. A partir del análisis realizado en la sección 2.4 del estado del arte, se han podido identificar un conjunto de recursos semánticos que ostentan ser los más utilizados en el área del PLN. Varios autores han creado recursos integradores, pero en su mayoría favoreciendo la composición lingüística más que la semántica (o conceptual). Surgiendo una interrogante. ¿Por qué si los recursos semánticos son tan usados y valiosos, existen tan pocas herramientas que proponen integrarlos y además no a la gran mayoría, si investigaciones han demostrado que la integración semántica en muchos casos, mejora los resultados de sistemas de PLN? En consecuencia de ello, en este capítulo se ha propuesto la elaboración de una base de conocimiento multidimensional que contenga la mayor cantidad de recursos semánticos integrados en una misma red de conocimiento.

Para su desarrollo se han propuesto dos fases de integración. La primera fase plantea fusionar cuatro recursos semánticos (ej. WN, WND, WNA, SUMO) y en la segunda se introducen algunas restructuraciones interrelacionales y la adición de dos conocimientos más (ej. SC y SWN). En general se centra en interrelacionarlos todos manteniendo como núcleo a WN (variando versiones ej. 1.6 y 2.0).

[67] http://lml.bas.bg/ranlp2011/
[68] http://gplsi.dlsi.ua.es/congresos/nldb11/
[69] http://gplsi.dlsi.ua.es/congresos/wassa2011/
[70] http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html

En el proceso de evaluación progresivo es posible detectar que con el uso de diferentes versiones de WN se pierden lazos, entonces, se propone en la segunda fase, la menor intervención posible de versiones intermedias para reducir pérdidas de enlaces. Luego del diseño de la estructura y las clases de implementación intervinientes, se obtiene una herramienta informática capaz de ser leída por sistemas del PLN con características semánticamente multidimensionales.

La herramienta, ha sido evaluada en ambas fases de desarrollo, para la primera se han elegido muestras representativas, donde se valora en qué porciento los elementos recuperados por el ISR-WN coinciden o recuperan la etiqueta o el *synset* asociado. Los resultados para esta fase están por encima del 94% de fiabilidad y un 100% con respecto al recurso SUMO. Como con SUMO no existen pérdidas, para la siguiente versión se procura interconectar *synsets* anotados en versiones iguales mientras sea posible. En esta fase, los resultados según la versión de WN utilizada como núcleo, han estado para WND y SUMO al 100% de fiabilidad y SC con 99,76% al 100% según WN1.6 y 2.0 respectivamente. Esto indica que la estrategia de reducción de versiones intermedias minimiza las pérdidas de enlaces. Respecto a WNA se aprecian algunas variaciones en las fiabilidades, pues se ha fusionado y complejizado. No obstante se considera confiable, pues todas las relaciones que presentan son exactas aunque algunas no se puedan establecer hasta tanto se remplace WNA por una versión más completa.

Una de las informaciones más distintivas introducidas en ISR-WN es SWN. Ahora no solamente se obtienen conceptos asociados a sentidos de palabras, sino que es posible conocer que tan positivo puede ser un sentido, dominio, categoría, emoción o clase semántica. A pesar que con la variación de versiones de anotación se desperdician descripciones de SWN, todos los enlaces analizados constatan su precisión al 100%.

Es importante resaltar que este recurso ha sido utilizado en varias investigaciones científicas reconocidas. En general, aprovechando la multidimensionalidad semántica que provee.

4. APROXIMACIONES DE RESOLUCIÓN DE AMBIGÜEDAD SEMÁNTICA BASADAS EN EL ANÁLISIS SEMÁNTICO MULTIDIMENSIONAL

En este capítulo se describen varias propuestas de resolución de ambigüedad semántica, que se consideran no supervisadas y basadas en conocimiento. Estas se dividen en dos grupos, Basadas en Árboles Semánticos y el segundo grupo, Basados en Grafos de Conocimiento. Dentro del primero, se presentan dos propuestas aptas para generar a partir de un texto, árboles semánticos con la capacidad de conceptualizar frases según la base de conocimiento utilizada. Con ese fin, se aplican medidas de conceptualización basadas en múltiples recursos. Además se introduce otra variante que, utiliza el mismo principio pero adicionando valores de frecuencia de sentidos en la medida de clasificación.

Dentro del grupo de propuestas basadas en grafos, se introduce por primera vez la visión del modelo de *N-Cliques* en WSD, con dos métodos. Además se plantea una nueva variante del uso del algoritmo *PageRank,* ahora combinado con frecuencia de sentidos. Todas las aproximaciones incluidas en ambos grupos se centran en medir la influencia de diferentes dimensiones semánticas en métodos de WSD. Para ello, se proponen un conjunto de evaluaciones comparativas minuciosamente detalladas, con el fin de conocer en qué medida se han superado o reducido los resultados alcanzados. Se tienen en consideración descripciones y reportes emitidos por otros investigadores.

4.1. INTRODUCCIÓN

En la actualidad, se han hecho populares diferentes aproximaciones que aplican métodos basados en grafos de conocimiento obteniendo resultados alentadores. Entre las diferentes técnicas utilizadas se pueden mencionar las aproximaciones que utilizan interconexiones estructurales, tales como SSI (Navigli and Velardi, 2005, Navigli and Velardi, 2004), capaz de crear características estructurales de los sentidos de cada palabra en un contexto mediante cadenas léxicas; otro trabajo relevante está relacionado con la exploración de la LKB generada por integración de WN y FrameNet (Laparra *et al.,* 2010) y por último, también cabe señalar aquellos basados en la LKB de WN (incluyendo las glosas desambiguadas) donde se aplica la muy conocida técnica de *PageRank* (Sinha and Mihalcea, 2007, Agirre and Soroa, 2009).

El uso de este tipo de aproximaciones tiene asociado el desarrollo de múltiples recursos semánticos. Muchos de estos están basados en WN, ya que este ha sido aceptado por la comunidad científica como el inventario de sentidos de referencia para la evaluación de diferentes sistemas de WSD.

Varios autores descritos en la sección 2.4.2.1 como (Magnini and Cavaglia, 2000), (Niles and Pease, 2001), (Niles and Pease, 2003), (Valitutti *et al.,* 2004) ,(Sara and Daniele, 2009), (Moldovan and Rus, 2001) y otros, han propuesto que se incorporen a la red semántica de WN, algunas taxonomías que caracterizan, en uno o varios conceptos a los sentidos de cada palabra (véase la sección 2.4.2). A pesar del hecho de que se han desarrollado una gran cantidad de nuevos recursos basados en WN, pocos logran integrar y relacionar la información semántica en un enfoque único (véase la sección 2.4.3). Entre los que más información semántica contienen integrada está MCR (Atserias *et al.,* 2004). Pero este recurso carece de algunas dimensiones de interés en esta Tesis como son: las SC (Izquierdo *et al.,* 2007), WNA (Strapparava and Valitutti, 2004) y para su uso en el próximo capítulo SWN (Esuli and Sebastiani, 2006). La ausencia de un recurso capaz de integrar y relacionar toda esta información semántica derivó en la creación del recurso ISR-WN como base de conocimiento principal de todas las aproximaciones de esta Tesis.

Es importante destacar que con el *baseline* MFS (basando en la frecuencia de sentidos), se ha conseguido alcanzar las primeras posiciones de los *rankings* de Senseval. Por ejemplo, en Senseval-2 un sistema utilizando el MFS ocuparía el segundo puesto con un 64,58% de *Precision* y *Recall* (Preiss, 2006); en Senseval-3 Denys Yuret de la Universidad de Koc calculó un 60,9% para ambas medidas y para la misma competición Bart Decadt de la Universidad de *Antwerp* anotó un 62,4%, estos resultados podrían situar los *baselines* en las posiciones séptima y quinta, respectivamente (Snyder and Palmer, 2004); en Semeval-1 el MFS se colocó en el noveno puesto entre catorce sistemas y para la competición Semeval-2 se ubicaría en un sexto lugar. Como se aprecia, este método probabilístico consigue obtener resultados eficaces en la tarea de WSD, sin embargo, no toma en cuenta la información contextual de ningún tipo, es decir, no constituye un método semántico real.

Todas estas valoraciones han provocado que en esta Tesis se aborden nuevas propuestas de WSD para aplicar Análisis Semántico Multidimensional en la tarea de Resolución de la Ambigüedad Semántica de las Palabras. Al tomar como referencia los resultados relevantes obtenidos por el *baseline* MFS, se ha optado por utilizar la frecuencia de sentidos también como una nueva dimensión. El objetivo es generar aproximaciones de WSD no supervisadas y basadas en conocimiento, capaces de superar otras del mismo tipo que aplican menos dimensiones. Para lograr ese propósito las aproximaciones de WSD presentadas en esta Tesis se muestran en el siguiente orden:

- Aproximaciones Basadas en Árboles
  - Árboles Semánticos Relevantes
  - Árboles Semánticos Relevantes combinados con Frecuencias de Sentidos
- Aproximaciones Basadas en Grafos
  - *N-Cliques* combinado con *Reuters Vector*
  - *N-Cliques* combinado con Árboles Semánticos Relevantes
  - *PageRank* combinado con Frecuencias de Sentidos

4.2. APROXIMACIONES BASADAS EN ÁRBOLES SEMÁNTICOS

En el proceso de desambiguación del sentido de las palabras se considera válido el uso de técnicas capaces de extraer características relevantes de los textos, para luego emitir criterios sobre estas y consecuentemente juzgar el texto asociado. Entre las características más populares están los términos relevantes presentes en un texto y los conceptos semánticos relevantes. Estos últimos, se identifican a partir de la información de corpus asociados a dominios (o categorías) o mediante recursos semánticos (redes semánticas). A continuación se introducen algunas aproximaciones de WSD, que hacen uso de la extracción de características conceptuales basadas en redes semánticas.

4.2.1. ÁRBOLES SEMÁNTICOS RELEVANTES

La propuesta de desambiguación mediante el uso de Árboles Semánticos Relevantes (*Relevant Semantic Trees* (RST) (Gutiérrez *et al.*, 2010b) identificados en los textos del lenguaje humano, consiste en los siguientes pasos:

1. Pre-procesamiento de corpus a analizar aplicando una herramienta *Pos-Tagger* (en este caso *Freeling* (Atserias *et al.*, 2006)) para obtener los lemas de cada palabra, categorías gramaticales y propuestas de frecuencias de sentidos.
2. Obtención de los Árboles Semánticos Relevantes para cada conjunto de lemas que representan a cada frase.
3. Selección de los sentidos correctos, siendo estos los más relacionados con los Árboles Semánticos Relevantes obtenidos.

Téngase en consideración, que estos pasos se ejecutan en una cadena de eventos donde la salida de uno constituye la entrada del próximo. A continuación se describe cada paso en detalle.

4.2.1.1. PRE-PROCESAMIENTO DE CORPUS

Se toma la frase de entrada que se introduce en *Freeling* para obtener las palabras que la componen. Según *Freeling,* por cada palabra se obtiene su categoría gramatical y una lista de *offsets* de WN ordenados por frecuencia (según su base de datos). En la Tabla 16 se detalla la salida de esta herramienta para la frase tomada del Corpus de la competición Senseval-2: *"The art of change ringing is peculiar to the English and like most English peculiarities unintelligible to the rest of the world".* El formato de la salida se corresponde con [palabra| lema| parte del habla (categoría gramatical)|lista de *offsets* …].

The the DT 0.9998
art art NN 1 00598038:02213100:04361152:05251961
of of IN 0.999901
change change NN 0.62963 00125689:02421790:02421923:03733220:05441797:07767320:09641836:09642046:09642453:09984639
ringing ringing NN 0.45 03898569:05397646:05502523
is be VBZ 0.999707 01552250:01666138:01775163:01775973:01781222:01782836:01784339:01787769:01811792:01817610:01840295:01843641
peculiar peculiar JJ 1 00347275:00458241:00919283:01054295
to to TO 0.999866
the the DT 0.9998
English english JJ 0.90625
and and CC 0.999723
like like JJ 0.0415267 01353127:01354252:01354638:01985976
most RBS RBS 0.998447
English english JJ 0.90625
peculiarities peculiarity NNS 1 02535416:03741160:04505614
unintelligible unintelligible JJ 1 00500611:01283982
to to TO 0.999866
the the DT 0.9998
rest rest NN 0.925926 00688597:03233549:05137171:09947291:10060454:10093091:10972097
of of IN 0.999901
the the DT 0.9998
world world NN 1 04384026:04476879:05957670:05973080:06079949:06684818:06691078:06753779

Tabla 16 Análisis léxico según *Freeling*.

Como se observa en la Tabla 16 *Freeling* puede producir diferentes errores:

1. Que la categoría gramatical que presenta una palabra en el contexto sea errónea (ej. *English* lo identifica como adjetivo y es un sustantivo).
2. Que se esté haciendo uso de una multi-palabra en la frase y no se detecte, como es el caso de *change- ringing*.
3. Que no detecte una contracción como por ejemplo Dr. y divida la frase en dos, lo que afectaría el análisis sintáctico de ambas oraciones.

El hecho de que *Freeling* se base en diccionarios deja una brecha abierta para los errores dos y tres, debido a que se limita al número de ejemplos que contiene. A continuación se ilustra un gráfico que describe la secuencia del proceso que se ejecuta con *Freeling*.

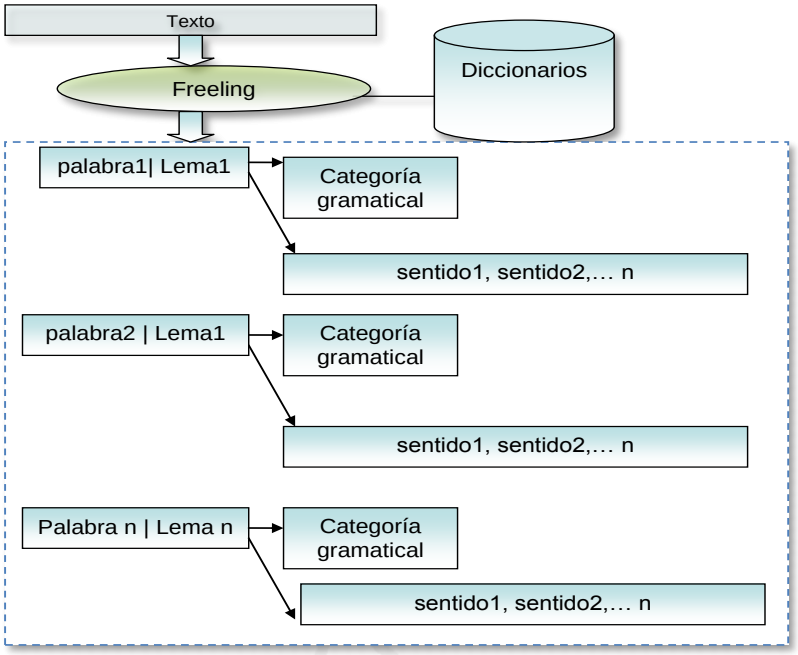


Figura 18. Secuencia del análisis léxico de *Freeling*.

4.2.1.2. PROCESO DE OBTENCIÓN DE CONCEPTOS RELEVANTES

Tras obtener los lemas de las palabras que intervienen en la frase, se utiliza el recurso ISR-WN para extraer de cada lema sus respectivos sentidos, que a su vez se encuentran asociados con conceptos de las tres jerarquías que interesan en esta propuesta (WND, WNA y SUMO). También se toma como jerarquía conceptual los hiperónimos relacionados con todos los sentidos del lema, tal y como lo hace (Montoyo, 2002), ya que los conceptos de dicha jerarquía caracterizan a sus descendientes. De WN solamente se trabaja con las relaciones taxonómicas, tomadas como un subconjunto de la red semántica. El objetivo final es obtener vectores de conceptos de las cuatro jerarquías que estén relacionados con la frase de forma global, y luego, de cada palabra, se elige el sentido que más conceptos en común tenga con los vectores. Para la construcción de estos vectores es necesario acumular los valores del cálculo de Ratio de Asociación (*Association Ratio* (AR)) y establecer las medidas de similitud. Todos estos pasos se describen a más adelante.

Con el objetivo de formalizar el tratamiento se tiene que a partir de una frase f, se extrae la lista de palabras $L = \{w^1, w^2, \dots, w^l\}$, las cuales se toman como entrada para el diccionario de WN con sus correspondientes representaciones en la base de conocimiento (LKB). Esto se traduce en WN $\subseteq$ LKB (por ejemplo, tomando como LKB al recurso ISR-WN). De esta manera, a cada palabra w^k se le asocia el conjunto de sentidos $sw^k = \{sw_1^k, sw_2^k, \dots, sw_n^k\}$, donde sw^k_j ($1 \leq j \leq n$) representa el j-ésimo sentido de la palabra w^k. Se comprende entonces la LKB por un grafo completo no dirigido $G = (V, E)$. En este grafo los vértices pueden estar representados por conceptos $C = \{c_1, c_2, \dots, c_r\}$ o por sentidos $s = \{s_1, s_2, \dots, s_m\}$, o sea $V = C \cup S$. La relación entre dos vértices v_i,v_j se representan como una arista $e_{i,j}$. Entonces, a partir de una lista L de palabras es posible obtener un conjunto de elementos semánticos $CS' = C' \cup sw$ siendo C' conceptos conformados por aquellos vértices $v_i \in C$, para los cuales existen aristas e_{ij}

donde $v_j \in sw^k$, tal que los sentidos sw^k se corresponden con las palabras incluidas en L siendo $sw^k \subseteq S$.

La propuesta de obtención de vectores de conceptos relevantes, difiere de (Vázquez *et al.*, 2004a) en que se tienen en cuenta cuatro jerarquías al mismo tiempo, además de aplicar una variación a la medida de AR. En la propuesta de Vázquez *et al.* se define w como una palabra (la cual se corresponde con w^k de la formalización presentada en esta Tesis) y D se define como un dominio (en la propuesta de esta Tesis se corresponde con el $v_i \in CS'$). El término D referido a los dominios ahora hace referencia a los conceptos de las taxonomías, incluso se pueden tomar los sentidos que representan a w (de este modo se valorarían los sentidos desde la perspectiva de Conceptos). La fórmula inicial para ser aplicada a las redes semánticas es la siguiente:

$$AR(w,D) = P(w,D) * \ log_2(\frac{P(w,D)}{P(w)}) \tag{31}$$

En el caso de la ecuación (31) se mide la relación de una palabra respecto a un dominio. Sin embargo, ahora lo que se desea analizar en esta Tesis es cuánto se asocia el dominio a la frase (lista de palabras). Por ese motivo, la ecuación anterior se interpreta de otra manera, proponiendo la siguiente ecuación (32).

$$AR(D,f) = P(D,f) * \ log_2(\frac{P(D,f)}{P(D)}) \tag{32}$$

Donde D= Concepto; f = frase (colección de palabras); $P(D,f)$ representa la probalidad de un dominio respecto a la frase; $P(D)$ representa la probalidad marginal del dominio.

Con esta propuesta se mide cuánto se asocia un concepto a la frase, similar a *Reuters Vector* (Magnini *et al.*, 2002), pero utilizando una ecuación diferente. En RST todos los conceptos asociados de las cuatro jerarquías implicadas cuantifican sus cercanías mediante el AR. Otras propuestas como la Densidad Conceptual (Agirre 1996) y del método *DRelevant* (Vázquez Pérez *et al.*, 2004), también obtienen los conceptos de una jerarquía que se asocian con una frase para adquirir información conceptual útil en el proceso de WSD. Para determinar el *AR* de un concepto (dominio) respecto a la frase, se aplica la ecuación (33).

$$AR(D,f) = \sum_{i=1}^{n} AR(D,f_i) \ ; \ \text{siendo} \ \ AR(D,w) = P(D,w) * \ log_2(\frac{P(D,w)}{P(D)}) \tag{33}$$

Donde f=conjunto de palabras (w); f_i=i-ésima palabra de la frase f

A continuación se describe la secuencia de creación de los vectores tomando como ejemplo una frase del corpus de Senseval-2.

Para la frase: *But it is unfair to dump on teachers as distinct from the educational establishment*.

Mediante el proceso ya analizado en la Figura 18 se obtiene la siguiente colección de lemas.

lemas $\{be, unfair;\ dump;\ teacher, distinct, educational, establishment\}$

Cada lema es localizado en el recurso ISR-WN y se asocia con los conceptos de SUMO, WND, WNA y WN. Tras aplicar la fórmula de *AR* (*D, f*) se obtienen colecciones en forma de vectores de conceptos para cada recurso. Los vectores resultantes siguen la siguiente estructura.

$$VARs = \{valor\ AR\_1 \mid concepto\_1\ ;\ valor\ AR\_2 \mid concepto\_2; \ldots valor\ AR\_n \mid concepto\_n\}$$

Donde $VARs$ indica vector *AR* de conceptos de SUMO y valor *AR* es el valor calculado al concepto respecto a la frase. El mismo método se hace para los restantes vectores a obtener: $VARd$ (vector *AR* de dominios), $VARa$ (vector *AR* de conceptos afectivos) y $VARwn$ (vector *AR*

de *synsets* de WN). En la Tabla 17 se muestran los resultados de *AR* (*D*, *f*) para cada uno de los recursos implicados.

Para el caso del vector de WN se coloca la primera palabra del conjunto de sinónimos, pero en realidad el identificador es el *offset* del *synset*, por eso se observa que se repite ***drop*** dos veces, lo que indica que son dos sentidos distintos para una misma palabra.

VARd		***VARs***		***VARa***		***VARwn***	
AR	*Domains*	*AR*	SUMO	*AR*	*Affects*	*AR*	WordNet
0.90	*Pedagogy*	1.68	*SubjectiveAssessmentAttribute*	8.25	*trait*	1.85	*organization*
0.90	*Administration*	0.90	*EducationalProcess*			0.86	*beginning*
0.36	*Buildings*	0.54	*Organization*			0.86	*abstraction*
0.36	*Politics*	0.54	*Motion*			0.86	*educator*
0.36	*Environment*	0.54	*Removing*			0.86	*structure*
0.36	*Commerce*	0.22	*Process*			0.86	*natural_process*
0.36	*Quality*	0.22	*Reasoning*			0.86	*body*
0.36	*Psychoanalysis*	0.22	*Government*			0.86	*proof*
0.36	*Economy*	0.22	*Creation*			0.86	*sell*
		0.22	*Selling*			0.86	*beat*
		0.22	*Impacting*			0.86	*get_rid_of*
		0.22	*OccupationalRole*			0.86	*discard*
						0.86	*drop#1*
						0.86	*drop#2*

Tabla 17. Vector inicial de conceptos de la frase.

4.2.1.2.1. CREACIÓN DE ÁRBOLES SEMÁNTICOS RELEVANTES

Mediante la ecuación presentada en la sección anterior se obtienen cuatro vectores. Para añadir más información semántica a los vectores anteriores se añade una premisa. "Si un concepto se asocia a la frase, entonces su hiperónimo también se asocia, pero con un valor menor de AR" (Gutiérrez et al., 2010b). La determinación de si se acepta agregar o no un concepto padre (hiperónimo), se hace de la siguiente forma.

Si $AR(PC, f) > 0$ se agrega al vector, donde el cálculo de AR de un concepto padre se hace según indica la ecuación (34):

$$AR(PC, f) = AR(ChC, f) - ND(IC, PC) \quad (34)$$

Donde ND (*Normalize Distance* (Distancia Normalizada)) se calcula según la ecuación (36).

$$ND(IC, PC) = \frac{Camino\_Mínimo(IC, PC)}{profundidad\_taxonomía} \quad (35)$$

Donde:

- PC: concepto padre para cada iteración (*Parent Concept*).
- ChC: concepto hijo en cada iteración (*Child Concept*).
- IC: es el concepto inicial del que se quieren agregar los ancestros (*Initial Concept*).
- $AR(ChC, f)$ obtiene el valor AR para un concepto ChC respecto a la frase, en caso que ChC forme parte del vector inicial de conceptos, $AR(ChC, f)$ se obtiene utilizando la ecuación (33), en caso contrario se debe obtener en una iteración anterior de la ecuación (34).

- $profundidad\_taxonomía$: profundidad del árbol jerárquico del recurso a utilizar.
- $Camino\_Mínimo$: obtiene un valor mínimo de tránsito entre dos conceptos mediante la relación de hiperonimia.

Tras haber agregado algunos padres (hiperónimos) hasta el máximo nivel posible, puede darse el caso de que exista duplicación de conceptos en los vectores. Para evitar esta situación, se reagrupan los conceptos y se acumulan los valores de AR de cada concepto. Entonces, se genera un vector final sin conceptos repetidos junto con los valores acumulados de $AR(D, f)$ según su presencia en el vector final. Para comprender mejor el método se puede observar la Figura 19, Figura 20 y Figura 21 que muestran un ejemplo de agregación de padres (hiperónimos) de algunos conceptos obtenidos en el vector $VARd$.

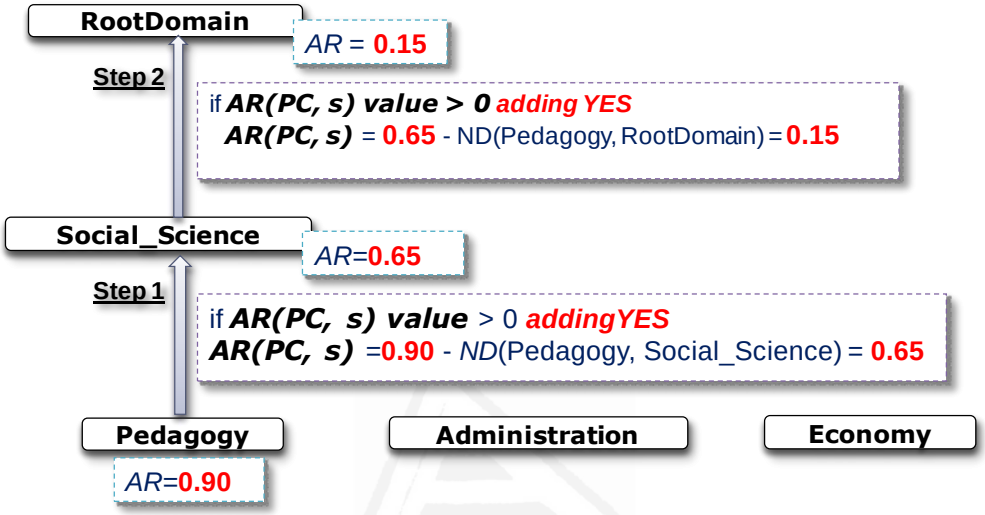


Figura 19. Primera iteración de agregar padres.

En la primera iteración (véase la Figura 19), se obtiene el valor de IC (ej. *Pedagogy*), después se aplica la ecuación (34) donde IC coincide con ChC y PC es *Social_Science*. El valor resultante de AR obtenido para *Social_Science* es de 0.65. Como consecuencia, al ser superior a cero se agrega este concepto al vector con ese valor. Para continuar agregando ancestros, se aplica la misma ecuación manteniendo IC = *Pedagogy* pero ChC ahora se corresponde con *Social_Science* y PC es *Root_Domain*. Finalmente se obtiene el correspondiente valor de AR para *Root_Domain*, agregándolo al vector. Un detalle importante es que a medida que se eleva el nivel de adición de padres dentro de cada iteración, ND irá aumentando, esto hace que a medida que se alejen de IC, se debilite su asociación con la frase.

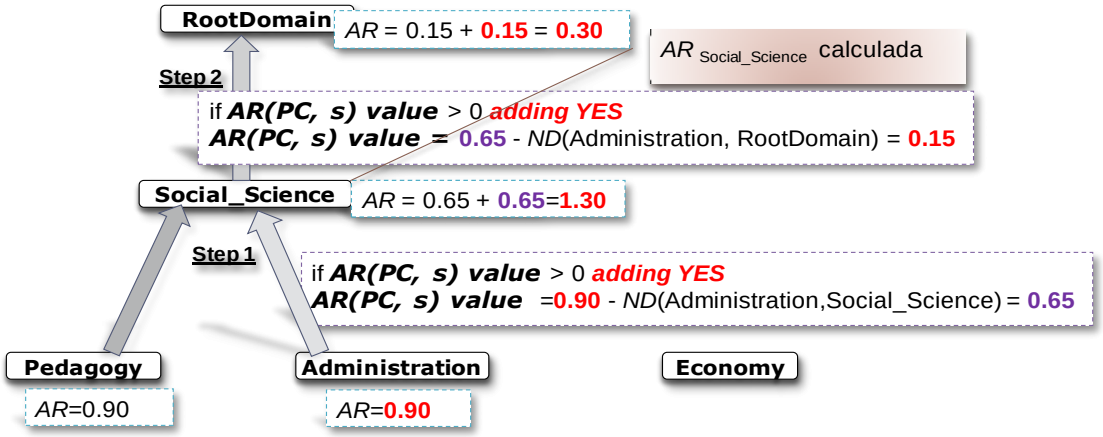


Figura 20. Segunda iteración de agregar padres.

En una segunda iteración (véase la Figura 20) para agregar los ancestros de *Administration*, ocurre que *Social_Science* ya existe en el vector. Para este caso lo que se hace es sumar el valor de *AR* obtenido por esta iteración, con el que tenía *Social_Science* previamente. Se debe prestar especial atención a que el valor utilizado por esta segunda iteración para calcular el *AR* de *Root_Domain* es el obtenido en esta iteración y no el que pueda haber acumulado. Finalmente, el proceso continúa de similar manera a la iteración anterior.

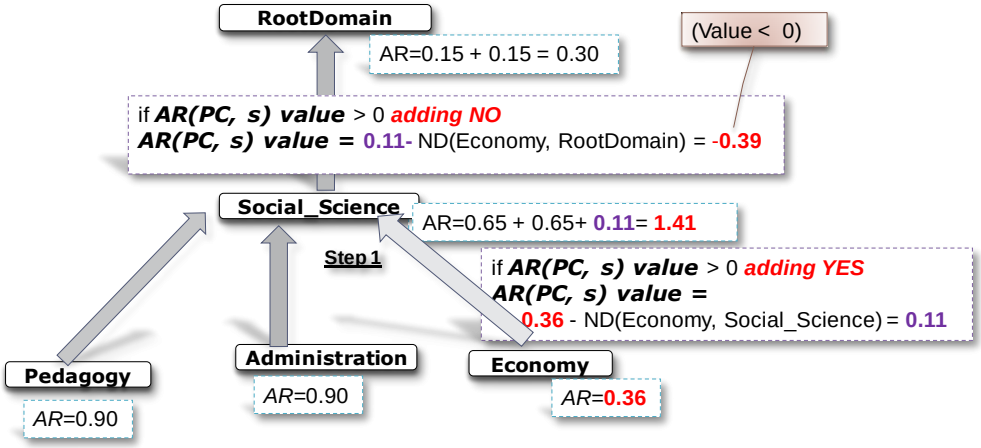


Figura 21. Tercera iteración de agregar padres.

Se puede apreciar en la tercera iteración (véase la Figura 21) que al intentar agregar a *Root_Domain*, sus conceptos descendientes no presentan valores de *AR* para ofrecer un resultado superior a cero. Entonces, el proceso de adición de conceptos padres en esa iteración termina ahí.

Todo el proceso tiene que cumplir que si $AR(PC, f) > 0$, se agrega el concepto padre con su valor de *AR* si no ha sido agregado previamente. Si ya ha sido agregado, se le suma el valor obtenido durante el proceso. Para el caso en que no se cumpla la condición, no se ejecuta el paso de agregación del concepto o suma del padre; esto sucede en la Figura 21, donde no se produce el segundo paso. Cabe destacar que el número de pasos está sujeto al valor de *AR* y a la profundidad en la que se encuentre el concepto de origen IC. Cada iteración para cada IC termina cuando el valor $AR \leq 0$ o no existen nuevos padres a agregar. El proceso de RST se aplica para cada elemento que esté presente en el vector inicial de Conceptos Relevantes, con el objetivo de construir un árbol semántico asociado a cada recurso implicado.

Después de haber aplicado el proceso de construcción de Árboles Semánticos Relevantes (conocido como *Relevant Semantic Trees* (RST) (Gutiérrez *et al.*, 2010b) se obtiene como resultado los elementos de la Tabla 18.

VARd		***VARs***		***VARa***		***VRAwn***	
AR	*Domains*	*AR*	SUMO	*AR*	*Affects*	*AR*	WordNet
1.63	*Social_Science*	2.68	*Entity*	8.25	*trait*	2.42	*social_group*
0.90	*Administration*	2.25	*RootSumo*	8.15	*root*	2.22	*group*
0.90	*Pedagogy*	1.81	*Process*			1.85	*organization*
0.80	*RootDomain*	1.72	*Physical*			1.62	*get_rid_of*
0.36	*Psychoanalysis*	1.69	*SubjectiveAssessmentAttribute*			1.54	*entity*
0.36	*Economy*	1.63	*RelationalAttribute*			1.48	*object*
0.36	*Quality*	1.61	*NormativeAttribute*			1.02	*cognition*
0.36	*Politics*	1.48	*Attribute*			0.86	*discard*
0.36	*Buildings*	1.40	*Abstract*			0.86	*abstraction*
0.36	*Commerce*	0.97	*Motion*			0.86	*sell*
0.36	*Environment*	0.91	*EducationalProcess*			0.86	*beginning*
0.11	*Factotum*	0.83	*OrganizationalProcess*			0.86	*natural_process*
0.11	*Psychology*	0.76	*IntentionalProcess*			0.86	*structure*

0.11	*Architecture*	0.63	*Organization*			0.86	*educator*
0.11	*Pure_Science*	0.56	*Transfer*			0.86	*body*
		0.55	*Removing*			0.86	*drop*
		0.49	*Group*			0.86	*beat*
		0.41	*Collection*			0.86	*drop*
		0.33	*Object*			0.86	*proof*
		0.23	*OccupationalRole*			0.82	*psychological_feature*
		0.23	*Reasoning*			0.76	*change_of_state*
		0.23	*Selling*			0.76	*move*
		0.23	*Creation*			0.76	*descend*
		0.23	*Government*			0.76	*evidence*
		0.23	*Impacting*			0.76	*gathering*
		0.16	*Touching*			0.76	*process*
		0.16	*FinancialTransaction*			0.76	*artifact*
		0.16	*InternalChange*			0.76	*professional*
		0.16	*PoliticalOrganization*			0.76	*exchange*
		0.16	*IntentionalPsychologicalProcess*			0.76	*concept*
		0.16	*SocialRole*			0.76	*strike*
		0.08	*Transaction*			0.66	*travel*
		0.08	*PsychologicalProcess*			0.66	*idea*
		0.01	*BiologicalProcess*			0.66	*touch*
		0.01	*ChangeOfPossession*			0.66	*information*
						0.66	*change*
						…..	…..

Tabla 18. Vector de Conceptos Final de la frase.

Se puede apreciar la diferencia de valores entre la Tabla 17 y Tabla 18, existe un incremento en la cantidad de conceptos de cada vector, identificando qué ramas de las jerarquías caracterizan más la frase. El árbol de la Figura 22 para el vector de WND ilustra con códigos de colores qué conceptos describen más la frase. Enmarcados mediante una elipse se destacan con colores más intensos los dominios de WND que acumulan los más altos valores de *AR* y además se encuentran agrupados bajo un mismo dominio. De esta forma, teniendo los cuatro vectores representados, se podrá determinar cuáles son los conceptos que caracterizan más la frase.

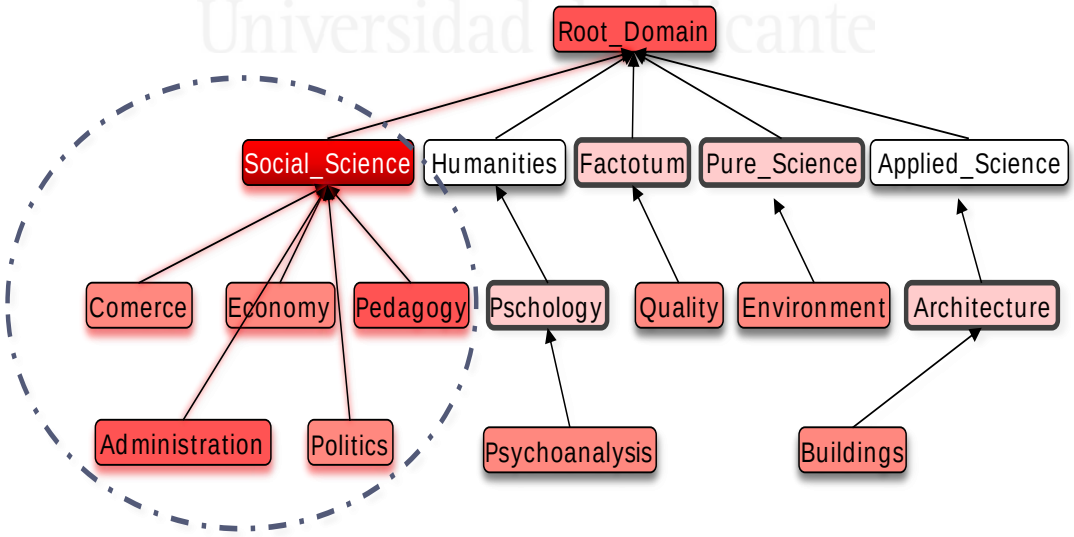


Figura 22. Árbol generado por el vector de WND.

4.2.1.3. SELECCIÓN DEL SENTIDO CORRECTO

Al poseer los patrones que caracterizan la frase, la siguiente etapa es determinar los sentidos correctos, para lo que se establecen los siguientes pasos:

1. De los posibles sentidos de la palabra a desambiguar, se deben discriminar los casos que no coincidan con la categoría gramatical que sugiere el *Pos-Tagger Freeling* en la sección 4.2.1.1. Para discriminar, se suma el valor uno al valor de $ACTs$ (Acumulado Total del sentido) del sentido que coincida en categoría y cero al que no. ¿Por qué no eliminar el que no coincida? Porque puede darse el caso de que *Freeling* se equivoque asignando la categoría gramatical de una palabra en la frase, pudiendo perderse el sentido correcto. Con esta idea se penalizan los sentidos menos posibles pero no se pierden (véase la Figura 23).

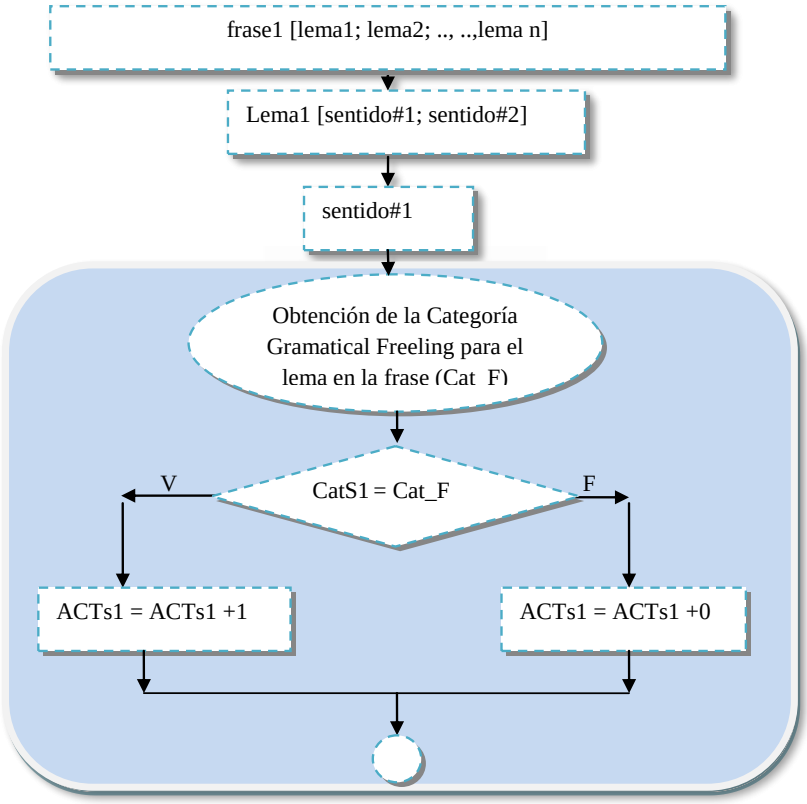


Figura 23. Discriminación por categoría gramatical.

- $ACTs1$: Acumulado total del sentido#1
- $CatS1$: Categoría del sentido#1
- Cat_F: Categoría de la palabra según *Freeling*

El objetivo del uso de $ACTs$ es para que cada sentido acumule el mayor valor de eventos positivos, como coincidir en categoría con *Freeling*, estar más identificado con los vectores y para casos de empate ser el de uso más frecuente. El proceso discriminatorio ejemplificado en la Figura 23 se le aplica a todos los sentidos de cada palabra.

En la Figura 24 se expone un diagrama de ejemplo con la palabra *unfair*.

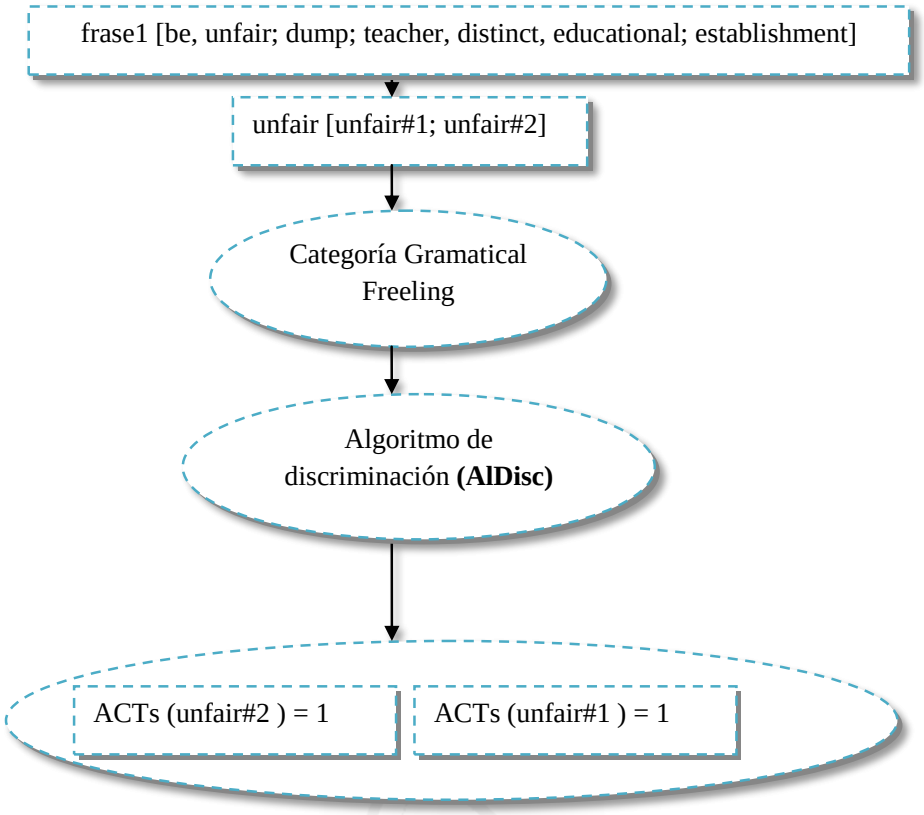


Figura 24. Discriminación de los sentidos de *unfair*.

Nótese que para este caso ambos sentidos coinciden con la categoría que el analizador léxico identifica, entonces se la ha agregado el valor de uno a ambos.

2. De cada sentido se obtienen RST's para cada recurso implicado (ej. los vectores Vss, Vds, Vas y $Vwns$) de igual forma que se ha aplicado para toda la frase, pero en este caso en particular únicamente a nivel de sentido y no a nivel de frase (véase el ejemplo de la Figura 25). Después por cada coocurrencia entre los vectores de la frase y los del sentido, se acumula el valor de AR del concepto del vector de la frase que coincida. La ecuación (36) es la responsable del proceso aplicada al vector de SUMO.

$$ACs(s, VARs) = \frac{\sum_k VARs[Vss^K]}{\sum_{i=1} VARs_i} \quad (36)$$

Donde:

- ACs: Acumulación de AR del sentido para el vector de AR de SUMO.
- $VARs$: Vector SUMO de conceptos relevantes de la frase, con formato $VARs$ [concepto | valor AR]
- Vss: Vector SUMO de conceptos relevantes del sentido, con formato Vss [concepto]
- Vss^k: es el k-ésimo concepto del vector Vss
- $VARs$ [Vss^k]: representa el valor de AR asignado al concepto Vss^k por el valor $VARs$.

El $VARs$ para el caso del sumatorio, accede a los valores de AR. La división de la acumulación de los $VARs$ coincidentes, entre el sumatorio de todos los AR de $VARs$, se aplica con el fin de normalizar los resultados. Este proceso se utiliza con las cuatro jerarquías según las ecuaciones (37), (38) y (39).

$$ACd(s, VARd) = \frac{\sum_k VARd[Vds^K]}{\sum_{i=1} VARd_i} \quad (37)$$

$$ACa(s, VARa) = \frac{\sum_k VARa[Vas^K]}{\sum_{i=1} VARa_i} \quad (38)$$

$$ACwn(s, VARwn) = \frac{\sum_k VARwn[Vwns^K]}{\sum_{i=1} VARwn_i} \quad (39)$$

Donde:

- ACd: Acumulación de AR del sentido para el vector de AR de WND.
- $VARd$: Vector WND de conceptos relevantes de la frase, con formato $VARd$ [concepto | valor AR].
- Vds: Vector WND de conceptos relevantes del sentido con formato Vds [concepto].
- Vds^k: es el k-ésimo concepto del vector *Vds.*
- $VARd$ [Vds^k]: representa el valor de *AR* asignado al concepto Vds^k por el valor $VARd$.
- ACa: Acumulación de AR del sentido para el vector de AR de WNA.
- $VARa$: Vector WNA de conceptos relevantes de la frase, con formato $VARa$ [concepto | valor AR].
- Vas: Vector WNA de conceptos relevantes del sentido, con formato *Vas* [concepto].
- Vas^k: es el *k*-ésimo concepto del vector Vas.
- $VARa$ [Vas^k]: representa el valor de *AR* asignado al concepto Vas^k por el valor $VARa$.
- $ACwn$: Acumulación de AR del sentido para el vector de AR de WN.
- $VARwn$: Vector WN de conceptos relevantes de la frase, con formato $VARwn$ [concepto | valor AR].
- $Vwns$: Vector WN de conceptos relevantes del sentido, con formato $Vwns$ [concepto].
- $Vwns^k$: es el *k*-ésimo concepto del vector $Vwns$.
- $VARwn$ [$Vwns^k$]: representa el valor de AR asignado al concepto $Vwns^k$ por el valor $VARwn$.

Al obtener los cuatro acumulados totales se procede al sumatorio de ellos almacenando el valor total de *AR* de conceptos coincidentes (véase la ecuación (40)).

$$ACTs = ACTs + \sum_{i=1} VARC_i \quad (40)$$

Donde:

- $ACTs$: Acumulado total de eventos positivos del sentido
- $VARC$=$\{ACs, ACd, ACa, ACwn\}$

Para entender de forma gráfica el proceso de acumulación se puede analizar la Figura 26. Aquí se expone un ejemplo donde al aplicar la ecuación sobre ACd se obtiene un valor AR acumulado, este proceso se hace extensivo para ACs, ACa *y* $ACwn$.

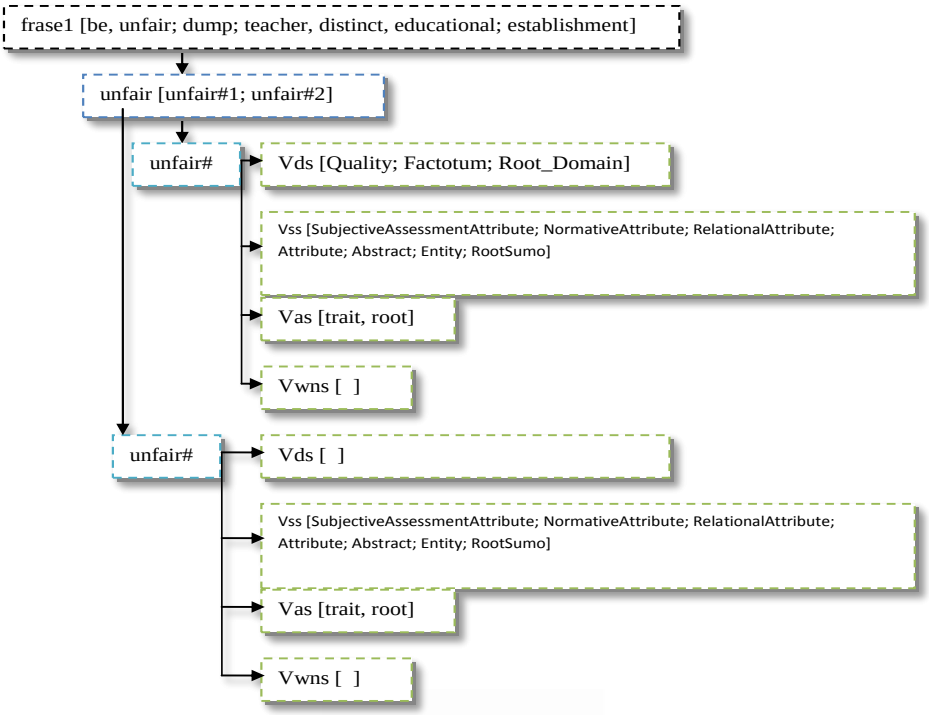


Figura 25. Vista de Vectores de los sentidos de *unfair*.

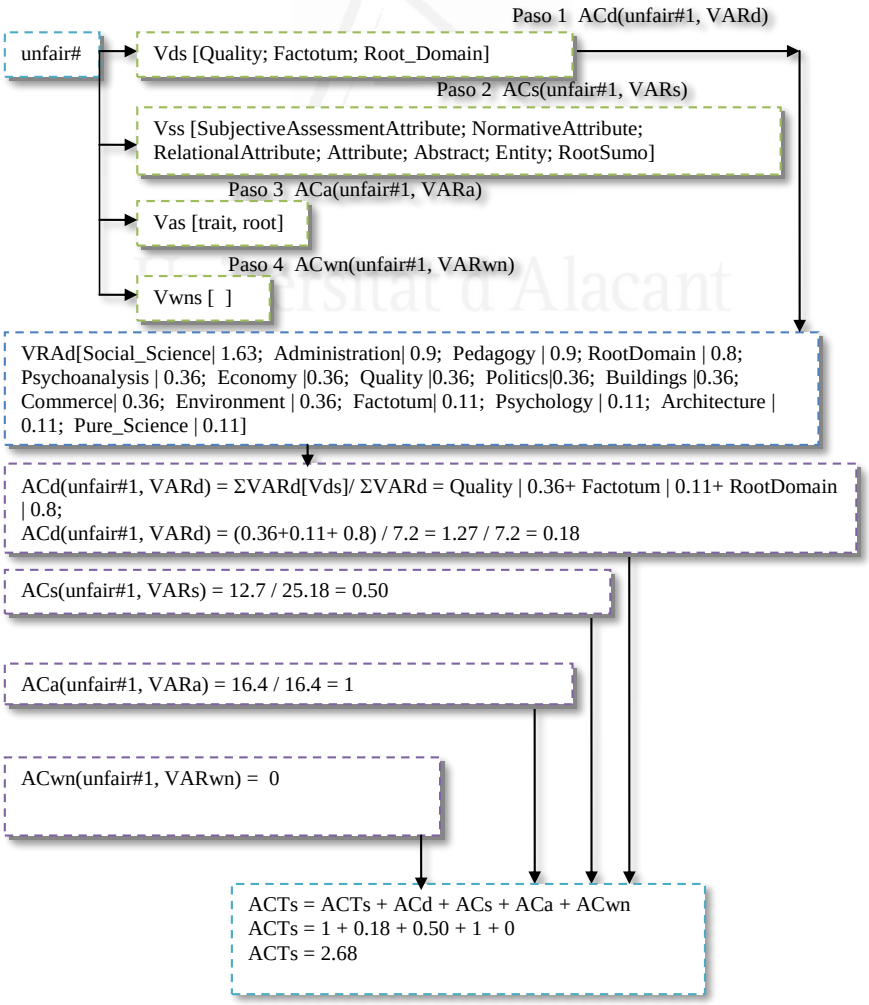


Figura 26. Proceso de *ACTs* para el sentido *unfair#1*.

El sentido que maximice el acumulado de relevancia es quien clasifica a la palabra objetivo. Si existe un empate de valor de $ACTs$ entre sentidos posibles, se les suma la frecuencia de sentidos que indica *Freeling* como valor de normalizado. Esto proporciona una la lista de sentidos ordenada por frecuencia y de esta forma se logra un desempate. La Figura 27 desarrolla el proceso de selección de sentidos.

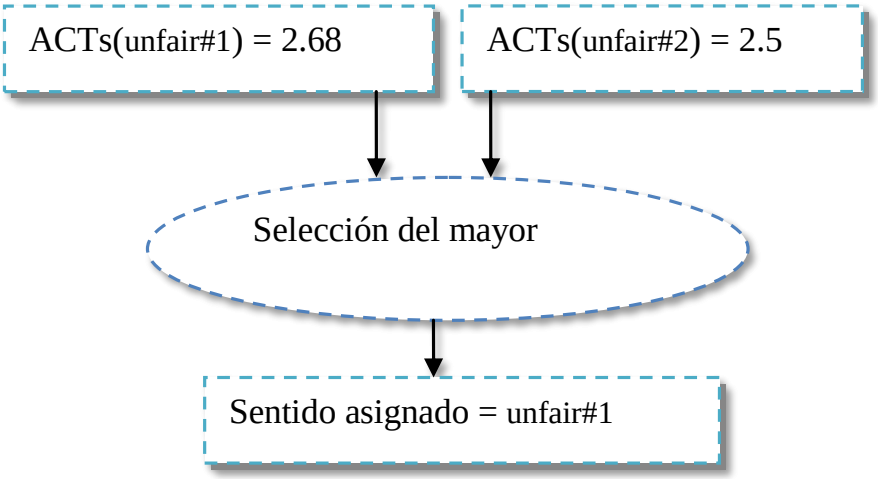


Figura 27. Selección del sentido correcto para *unfair*.

Para el caso hipotético que entre los valores de acumulado exista empate, se suma a $ACTs$ de cada sentido la frecuencia normalizada que se obtiene de *Freeling*. En la Figura 28 se muestra un caso hipotético con empates de $ACTs$. Donde FNF es la frecuencia normalizada de *Freeling*. Este valor corresponde al valor de frecuencia de cada sentido dividido entre el total de frecuencias de los sentidos de una misma palabra.

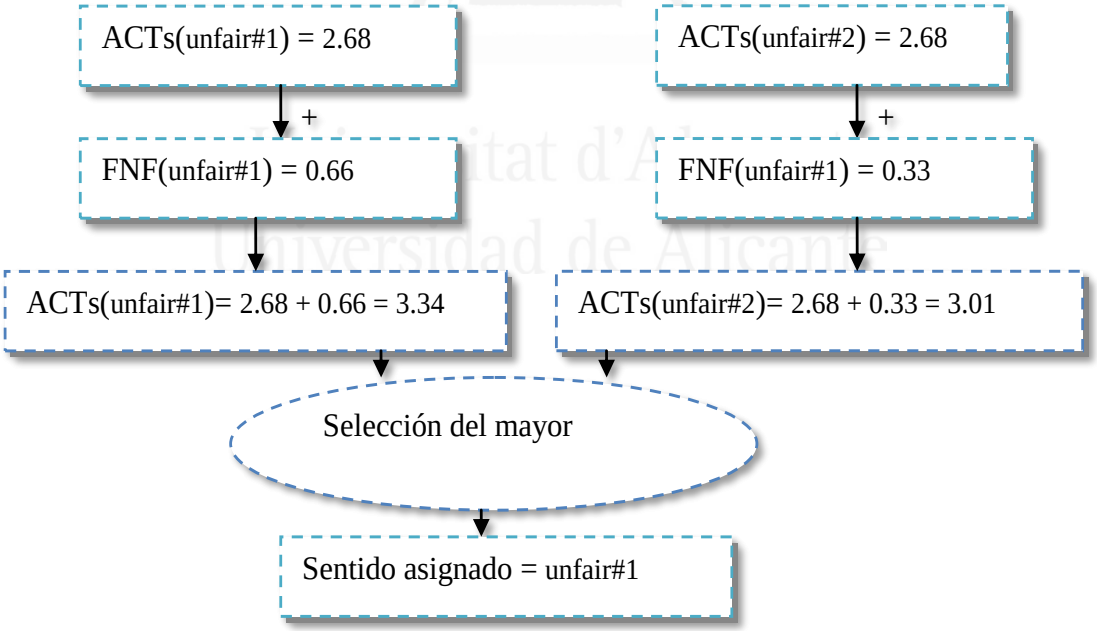


Figura 28. Proceso hipotético de discriminación por frecuencia de uso para *unfair*.

En la Tabla 19 se muestran para cada sentido posible de cada lema el valor normalizado de acumulación de AR, correspondiente al ejemplo del que se ha mostrado su desarrollo.

Lema	*Key*	*ACTs*
unfair	*unfair%3:00:00::*	2.68
	unfair%3:00:04::	2.5
dump	*dump%1:10:00::*	0.45
	dump%1:15:00::	0.49
	dump%1:22:00::	0.57
	dump%2:35:00::	1.46
	dump%2:38:01::	1.42
	dump%2:38:00::	1.41
	dump%2:40:02::	1.45
	dump%2:40:00::	1.47
	dump%2:40:01::	2.27
teacher	*teacher%1:09:00::*	1.54
	teacher%1:18:00::	2.06
distinct	*distinct%5:00:00:clear:00*	1.83
	distinct%5:00:00:definite:00	1.77
	distinct%3:00:00::	1.7
	distinct%5:00:00:different:00	1
	distinct%5:00:00:separate:00	1
educational	*educational%5:00:00:instructive:00*	1.43
	educational%3:01:00::	2.07
establishment	*establishment%1:04:00::*	1.43
	establishment%1:06:00::	1.71
	establishment%1:09:00::	1.46
	establishment%1:14:00::	1.95
	establishment%1:14:01::	1.15
	establishment%1:14:02::	2.35
	establishment%1:22:00::	1.56

Tabla 19. Sentidos posibles con su correspondiente $ACTs$.

Según el paso número tres, se selecciona de cada lista de sentidos correspondiente a cada lema, el de mayor $ACTs$. En la Tabla 20 se listan los sentidos seleccionados después de concluir el proceso.

Lemas	***Keys***
unfair	*unfair%3:00:00::*
dump	*dump%2:40:01::*
teacher	*teacher%1:18:00::*
distinct	*distinct%5:00:00:clear:00*
educational	*educational%3:01:00::*
establishment	*establishment%1:14:02::*

Tabla 20. *Keys* de los sentidos seleccionados.

En la Tabla 21 se establecen las coincidencias que se obtuvieron al aplicar el método propuesto, los *Keys-Senseval-2* corresponden con los identificadores que Senseval-2 anota como correctos para la tarea *English All Words*.

Lemas	*Keys* obtenidos	*Keys*-Senseval-2	*AR* Selecionado
Unfair	***unfair%3:00:00::***	***unfair%3:00:00::***	*2.68*
Dump	*dump%2:40:01::*	*dump%2:38:00::*	*2.21*
Teacher	***teacher%1:18:00::***	***teacher%1:18:00::***	*2.06*
Distinct	*distinct%5:00:00:clear:00*	*distinct%5:00:00:separate:00*	*1.83*
educational	***educational%3:01:00::***	***educational%3:01:00::***	*2.07*
establishment	*establishment%1:14:02::*	*establishment%1:14:00::*	*2.35*

Tabla 21. Coincidencias de sentidos con Senseval-2.

Tras analizar las coincidencias de este ejemplo se obtiene un 50% de *Precision* y *Recall*, e incluso después de haber estudiado la desambiguación del corpus de Senseval-2 completo, de los sentidos que fallan, algunos se encuentran cerca del mayor acumulador de AR. Es importante destacar que el método propuesto puede indistintamente asumir uno u otro vector en dependencia dependiendo de con qué recursos se desee medir WSD.

4.2.2. ÁRBOLES SEMÁNTICOS RELEVANTES COMBINADOS CON FRECUENCIAS DE SENTIDOS

Los RST's propuestos anteriormente proporcionan la conceptualización de frases. Este método de WSD puede mejorarse incluyendo algunos cambios. Tras analizar del comportamiento del *baseline* MFS en cada competición, surge una nueva alternativa, de la cual RST puede beneficiarse. Observando el *baseline*, se aprecia que siempre se ha posicionado entre los primeros lugares del *ranking* (véase la sección 2.6.1). Por ejemplo, analizando su *Precision* y *Recall* para *All Words*, en Senseval-2 obtiene un 64.58%, posicionándose en el segundo lugar. Luego, en Senseval-3 60.9% y 62.4% en dos sistemas diferentes correspondiéndoles los puestos séptimo y quinto respectivamente. En Semeval-1 logra un noveno puesto entre catorce sistemas y por último en Semeval-2 un sexto lugar. Estos resultados indican que este método probabilístico es capaz de obtener resultados efectivos sin tener en cuenta la semántica del texto. Teniendo en cuenta esos hechos, en esta sección se asume que es importante para la tarea de WSD poder obtener el sentido más frecuente asociado a la información contextual. Luego de estas consideraciones y motivado por ideas similares como las planteadas por (Mc.Carthy *et al.*, 2004), se pretende desarrollar un método informático no supervisado y basado en conocimiento que use la frecuencia de sentidos combinada con técnicas que tengan en consideración la semántica del texto, para ser capaz de superar a ambos por separado.

En concreto se propone utilizar la propuesta original de RST descrita en la sección anterior 4.2.1, incluyendo en sus ecuaciones la frecuencia de sentidos obtenida del análisis del corpus de SemCor. Resultando entonces, en un proceso de votación entre las salidas obtenidas desde la perspectiva de diferentes recursos. El proceso de votación involucra a MFS como recurso, y los resultados de RST combinados con Frecuencias de Sentidos ($\mathrm{RST} + \mathrm{Frec}$) (Gutiérrez *et al.*, 2011b) sobre WND, WNA, WN y SUMO. Esta idea específicamente se considera un tipo de ayuda supervisada (ej. MFS) a la propuesta no supervisada RST planteada anteriormente.

Consiste en dos fases principales aplicadas a cada palabra objetivo:

- Fase 1. Obtención de RST's.
- Fase 2. Selección de sentidos correctos:
    - Paso 1. Obtención de RST's de los sentidos candidatos.
    - Paso 2. Obtención de valores acumulados de relevancia para cada recurso y frecuencia de sentidos.
    - Paso 3. Proceso de votación para obtener el sentido final.

A continuación se presenta en detalle cómo se desarrollan cada una de las fases.

4.2.2.1. OBTENCIÓN DE LOS ÁRBOLES SEMÁNTICOS RELEVANTES

El proceso aplicado en este paso coincide totalmente con el descrito en las secciones 4.2.1.1 y 4.2.1.2.

4.2.2.2. SELECCIÓN DEL SENTIDO CORRECTO

En esta fase se introducen algunas modificaciones a la ecuación del cálculo de asociaciones acumuladas descritas en la sección 4.2.1.3. Además, se introduce una propuesta de votación en contraste con el sumatorio de valores de AR presentado en dicha sección. Para la selección del sentido correcto que aquí se defiende se aplican ahora tres pasos.

4.2.2.2.1. PASO 1. OBTENCIÓN DE RST DE LOS SENTIDOS CANDIDATOS

En este paso se persigue asociar cada sentido candidato de la palabra objetivo a un Árbol Semántico Relevante basado en cada dimensión semántica tenida en consideración. El objetivo es obtener para cada sentido los RST's, para luego establecer similitudes entre los $RSTs$ de la frase y los de cada sentido. Con relación a cumplir esa acción, se aplica un proceso similar al de la sección 4.2.1.3 utilizando la ecuación (33) donde se sustituye la variable w (palabra) por la variable sw^k_i, indicando el i-ésimo sentido de la palabra w^k. Los conceptos resultantes se almacenan en un vector de estructura [Concepto, valor AR]. Luego, se continúa aplicando el proceso de construcción del árbol descrito en la sección 4.2.1.3.

4.2.2.2.2. PASO 2. OBTENCIÓN DE VALORES ACUMULADOS DE RELEVANCIA PARA CADA RECURSO Y FRECUENCIA DE SENTIDOS

Para medir la similitud entre RST's de frase y sentidos, se introduce una modificación a las ecuaciones (36), (37), (38) y (39) introduciendo los valores de frecuencia de sentidos ($Frec_s$). Estas ecuaciones responden a un comportamiento común, es decir, todas son iguales, solamente cambia el vector del recurso a utilizar. El objetivo que se persigue es el de obtener un nuevo valor, para conseguir el sentido más frecuente en un contexto determinado. En la ecuación (41) el valor de *AR* del vector de la frase es acumulado (AC) cuando coinciden los elementos de los vectores de RST de frase y del sentido candidato.

$$AC(s, VAR) = \frac{\sum_k VAR[Vs^k]}{\sum_{i=1} VAR_i} + Frec_s \quad (41)$$

Donde:

- AC es el valor AR acumulado para los elementos analizados.
- VAR es el vector de conceptos relevantes de la frase, con formato: VAR [Concepto1 |valor $AR, \ldots$].
- Vs es el vector de conceptos relevantes del sentido candidato, con formato: Vs[Conceptos].
- Vs^k es el k-ésimo concepto del vector *Vs*.
- VAR [Vs^k] representa el valor de *AR* asignado al concepto Vs^k en ARV;
- $Frec_s$ representa el valor normalizado de frecuencia de sentidos para el sentido s obtenido del fichero *cntlist* de WN 1.6.

- $\sum_{i=1} VAR_i$ es el término que normaliza el resultado.

El valor de AC es calculado para cada RST (o Vector Relevante) de cada dimensión semántica. En esta propuesta se obtienen cuatro valores de AC (para la taxonomía de WN, WND, WNA y SUMO). Nótese que una vez obtenido estos valores para cada sentido de la palabra objetivo en cada dimensión, si el sentido no coincide con la categoría gramatical que *Freeling* sugiere (Atserias *et al.*, 2006), se aplica la misma discriminación descrita en la sección 4.2.1.3, añadiendo valor cero a AC en caso de no coincidir con *Freeling* y valor uno en caso contrario.

Finalmente el sentido candidato para cada recurso será en de mayor *AC* entre todos los sentidos de la palabra objetivo.

4.2.2.2.3. PASO 3. PROCESO DE VOTACIÓN PARA OBTENER EL SENTIDO FINAL

Como se ha explicado anteriormente, cada dimensión semántica provee un posible sentido en cada análisis. Es importante resaltar que en este paso de votaciones la frecuencia de sentidos se toma como una dimensión más. El proceso de votaciones involucra ahora a cinco dimensiones y en caso de no existir un acuerdo, el resultado final lo decidiría el MFS. Esta decisión es debido a que estudios empíricos han demostrado ser la mejor opción de desempate (Molina *et al.*, 2002). La ecuación (42) es la encargada de aplicar la votación.

$$Ps = max_i(\underset{k}{max}(V[VAC]_k)_i) \tag{42}$$

Donde:

- VAC corresponde al vector compuesto por valores de AC de cada sentido para una palabra objetivo.
- V es un vector de VAC.
- k corresponde a cada recurso o dimensión.
- $V\ [VAC]_k$ corresponde al k-ésimo VAC para el recurso k.
- $\max_k(V\ [VAC]_k)$ determina el sentido con el máximo valor de AC de cada VAC.
- i es el i-ésimo sentido.
- max_i determina el sentido que ha sido seleccionado en más ocasiones (max_k) entre todos los recursos.
- Ps: indica el sentido finalmente propuesto para una palabra objetivo.

El formato de VAC es: VAC [valor AC| sentido#1, valor AC| sentido#2, valor *AC*| sentido#n], y el formato de V [VAC] es: V [VAC-*Domains*, VAC-*Affects*, *VAC*- Taxonomías WordNet, VAC-SUMO, VAC-Frecuencias de sentidos]

4.3. APROXIMACIONES BASADAS EN GRAFOS

En esta sección se presentan dos aproximaciones para resolver la ambigüedad semántica basadas en técnicas de grafos. La comunidad científica desde muchos años atrás ha desarrollado diversos algoritmos y técnicas que se aplican a estructuras de redes (ej. (Dijkstra, 1959) (Floyd, 1963)), obteniendo diferentes resultados (ej. obtención de caminos mínimos, poda de estructuras, agrupamiento de elementos, detección de elementos relevantes, optimización de algoritmos, etc.). A medida que progresan otras áreas de investigación ajenas a la algoritmia y a la informática aplicada en sí, estas técnicas se han ido utilizando con fines más concretos y empleados en diferentes contextos científicos. Por ejemplo, la técnica de *PageRank* introducida por (Brin and Page, 1998) y ha sido utilizada por el buscador de Google[71] (Haveliwala, 2003) para asignar valores de relevancia a los sitios webs, o la técnica de *Cliques* introducida por (Luce and Perry, 1949) ha sido muy utilizada en las redes sociales para conocer los conjuntos de usuarios más relacionados entre sí.

Entre las áreas en las que se han utilizado técnicas basadas en grafos, se encuentra PLN, y en particular la tarea de WSD. Distintos sistemas basados en conocimiento han obtenido muy buenos resultados utilizando técnicas basadas en grafos (véase la sección 2.5.4.2). Se pueden mencionar los enfoques que utilizan las interconexiones estructurales, tales como SSI (Navigli and Velardi, 2005) que crean las especificaciones estructurales de los sentidos posibles de cada palabra en un contexto. Otro enfoque es la propuesta de explorar la integración de WordNet (WN) y FrameNet (Laparra *et al.*, 2010) y entre los más relevantes se pueden mencionar aquellos que usan *PageRank* como (Agirre and Soroa, 2009), (Reddy et al., 2010) (Soroa *et al.*, 2010), (Soroa *et al.*, 2010) y (Sinha and Mihalcea, 2007) que utilizan las interconexiones internas de la base de conocimiento léxico (*Lexical Knowledge Bases* (LKB) + *eXtended* WN) de WordNet. Debido a la gran popularidad que han adquirido las aproximaciones basadas en grafos y de acuerdo con sus exitosas demostraciones de efectividad, se han reducido las distancias que por mucho tiempo han existido entre sistemas supervisados y los que no lo son. En las siguientes secciones se presentan distintas propuestas de WSD basadas en redes semánticas, en concreto en ISR-WN y en WN+XWN que buscan mejorar los logros que se han alcanzado por autores que aplican este tipo de estructuras.

A continuación se describen dos grupos de aproximaciones de WSD basadas en redes semánticas. El primero describe dos aproximaciones basadas en la adaptación del modelo de *Cliques* para su uso en WSD y el segundo presenta una propuesta basada en *PageRank*.

4.3.1. APLICACIÓN DE TÉCNICAS DE *N-CLIQUES*

Los *Cliques* fueron definidos formalmente por (Luce and Perry, 1949) en términos de amistad de la siguiente forma: "Un *Clique* es un conjunto de más de dos personas si todos ellos son amigos mutuos el uno del otro". Con respecto a entender qué es un *Clique* es mejor seguir la explicación proporcionada por (Cavique *et al.*, 2009): "Teniendo un grafo no dirigido $G = (V, E)$ donde V denota el conjunto de vértices y E en conjunto de aristas, el grafo $G_1 = (V_1, E_1)$ es llamado un sub-grafo de G si $V_1 \subseteq V$, $E_1 \subseteq E$ y para toda arista $(v_i, v_j) \in E_1$ los vértices $v_i, v_j \in V_1$. Un sub-grafo G_1 se define como completo si en él existe una arista para cada par de vértices". Entonces un sub-grafo completo es también llamado *Clique*.

Para la obtención de conjuntos de elementos de una estructura de grafo, varios autores han trabajado en alternativas de *Cliques* o *cluster* de grafos. Se pueden mencionar (Luce, 1950,

[71] http://www.google.com

Balasundaram et al., 2006) con *N-Cliques*, (Balasundaram *et al.*, 2006) con *K-plex* y (Mokken, 1979) con *Clubs* y *Clans*. También se pueden encontrar varios autores que han analizado la aproximación de *Clique* como un problema NP-Completo (*Not Polynomial Complete*) (Wood, 1997). En este trabajo, para la resolución de la ambigüedad se ha seleccionado el modelo *N-Cliques* debido a que su topología se acerca a los requerimientos necesarios de la tarea WSD. Esto no significa que las demás propuestas no puedan ser aplicadas en un futuro.

4.3.1.1. N-CLIQUES

El modelo *N-Cliques* fue introducido por (Luce, 1950), donde define un *N-Clique* de un grafo *G* como un sub-grafo de *G* inducido por un conjunto de vértices *V* asociados a un sub-grafo máximo completo de potencia G^n. Con el fin de conocer el significado de *n* se sigue la definición del artículo de (Alba, 1973) que dice: "... vamos a suponer que el grafo *G* es conexo y tiene un diámetro estrictamente mayor que n. La n-ésima potencia de G^n de G es un grafo con $V(G^n) = V(G)$ y tal que v_i y v_j son adyacentes en G^n si y sólo si $dG(v_i, v_j) \leq n$, es decir, dos vértices son adyacentes en G^n cuando la distancia entre ellos en el *G* es n o menos. "

Decir 1-*Clique* es igual que referirse a *Clique*, porque la distancia entre los vértices es de una arista. 2-*Clique* es el sub-grafo máximo completo con longitud de trayectoria de una o dos aristas. Con la aplicación de un grafo de *N-Cliques* se pueden obtener diferentes sub-conjuntos fuertemente integrados (Friedkin, 1984). Estos modelos han sido aplicados a diferentes tipos de conexión de redes para fines muy distintos. Por ejemplo, se han utilizado para las redes de difusión (redes celulares) (Clark *et al.*, 1990), en procesos biomédicos (Kose *et al.*, 2001), en PLN para la adquisición léxica (Widdows and Dorow, 2002) y en otras muchas aplicaciones. Para utilizar el modelo *N-Cliques* en esta Tesis se ha propuesto una modificación del algoritmo heurístico original llamado Algoritmo de Particionamiento de *Clique* (*Clique Partitioning Algorithm*).

4.3.1.2. ALGORITMO DE PARTICIONAMIENTO DE *CLIQUES*

Este algoritmo fue presentado por (Tseng and Siewiorek, 1986) únicamente para el modelo de *1-Clique* o *Clique*. En este trabajo es necesario obtener relaciones entre elementos para distancias entre vértices de más de una arista, con lo cual, se introduce una pequeña modificación con el fin de aplicar este algoritmo con un comportamiento similar a *N-Cliques*. El algoritmo original se muestra en la Tabla 22, donde los cambios introducidos se han subrayado.

Tabla 22. Algoritmo de particionamiento de *N-Cliques.*

```
/*Create super graph G´(S,E) S: SuperNodes, E: edges */
/*V: vertex, N:distance between pair nodes*/
S = Ø; E' = Ø;
for each vi ∈ V do si = {vi}; S = S ∈{si}; endfor
for each ei,j ∈ E do E' = E' ∈{e'i,j}; endfor
while E'≠ Ø do
  /* Find sindex1, sindex2 having most common neighbors*/
  MostCommons = -1;
  for each e'i,j ∈ E' do
    ci,j = |COMMON-NEIGHBOR(G', si , sj, N )|;
    if ci,j > MostCommons then
      MostCommons = ci,j ;  Index1 = i;  Index2 = j;
    endif
  endfor
  CommonSet = COMMON-NEIGHBOR (G',sindex1 , sindex2, 1);
  /*delete all edges linking Sindex1 or Sindex2 */
```

E' = DELETE-EDGE(E', s_{Indexl}); E' = DELETE-EDGE(E', s_{Index2});
*/*Merge s_{index1} and s_{index2} into $S_{index1Index2}$ */*
$S_{index1Index2} = S_{index1} \cup S_{index2}$;
$S = S - S_{index\ 1} - S_{index\ 2}$; $S = S \cup \{S_{index\ 1Index\ 2}\}$;
/ add edge from $S_{IndexlIndex2}$ to super-nodes in CommonSet */*
for each $s_i \in$ CommonSet do
$E' = E' \cup \{e'_{i,Index\ 1Index\ 2}\}$;
Enfor s_i
/* Decrease in 1 the N value */
If N > 1 then N = N-1;
endif
Endwhile Return S;

Esta modificación es capaz de crear un *Clique* a distancia N y luego a distancias $N-1, N-2, \ldots N=1$ mientras existan aristas en la red. La función $COMMON-NEIGHBOR\ (G_1, s_i, s_j, N)$ devuelve el conjunto de super-nodos que tienen vecinos comunes a distancia N de cada s_i y s_j en G'. $DELETE-EDGE\ (E_1, s_i)$ elimina todos los lazos en E' los cuales tienen a s_i como su interfaz de super-nodo. Esto significa que al ser s_i elegido para formar parte de un nuevo super-nodo, sus lazos como nodo individual deben eliminarse y cederse al nuevo super-nodo, que conservará los lazos comunes entre s_i y de s_j). Cabe destacar que $CommonSet$ contiene el conjunto de super-nodos que son vecinos comunes a distancia de una arista de s_{index1} y s_{index2} en G'. Para comprender mejor la ejecución de este algoritmo en la Figura 29 se muestra un ejemplo usando como entrada $N = 2$ (distancia entre vértices de dos aristas como máximo).

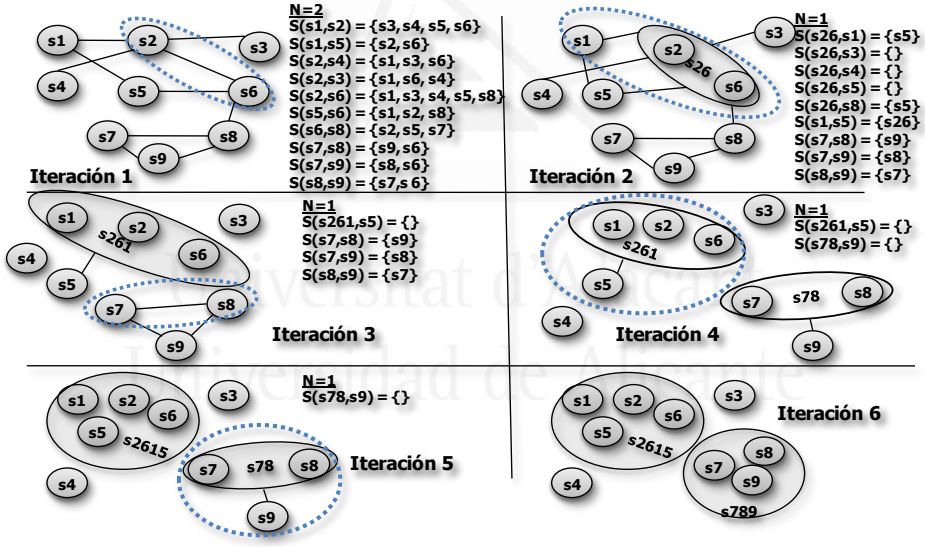


Figura 29. Ejemplo del algoritmo de particionamiento de *N-Clique* con $N = 2$.

Como se observa, este algoritmo heurístico es capaz de obtener un conjunto de nodos con la distancia máxima entre todos los nodos ≤ 2 aristas entre sí, ya que $N = 2$. Este algoritmo continúa con la obtención de los siguientes *Cliques*, reduciendo las distancias entre vértices en una arista $N = N - 1$, donde $N \geq 1$ para cada iteración, mientras que $E \neq \emptyset$.

A continuación se describen dos propuestas que aplican este particionamiento usando la técnica de *N-Clique*. Ambas utilizan la misma base de conocimiento, en este caso el recurso ISR-WN, pero varía el modo en que se construye la red que representa a cada frase.

Respecto al uso de ISR-WN, este recurso proporciona una herramienta que permite navegar desde cualquier sentido de WN, etiquetas de dominio o categorías de SUMO a través de las relaciones internas como se muestra en la Figura 14. Visto así, se puede descubrir la multidimensionalidad de conceptos que existe en cada frase. La Figura 30 muestra cómo los

conceptos caracterizan a una frase y cómo las palabras se relacionan a través de la red semántica.

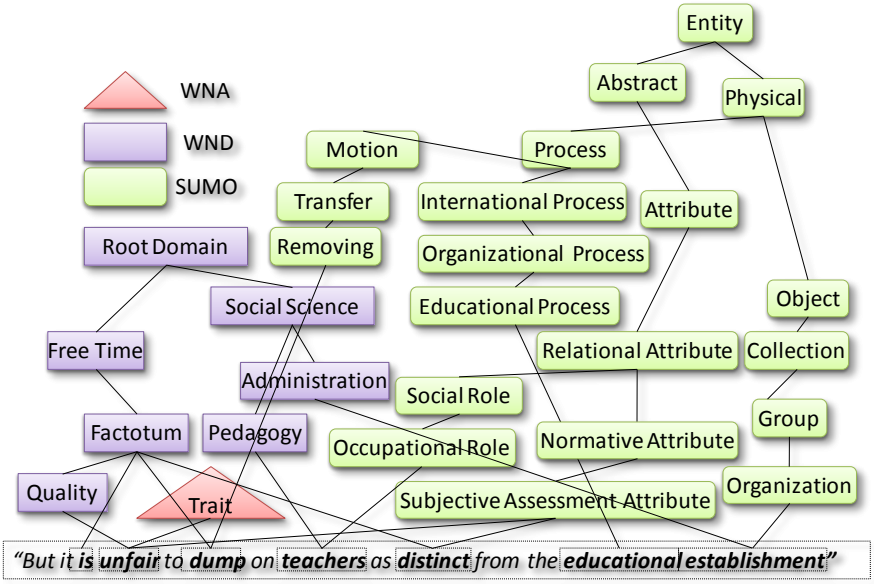


Figura 30. Extracción de características semánticas de la frase.

A continuación se describe la propuesta en cuestión (Gutiérrez et al., 2011d), detallando cada paso a seguir hasta la obtención final de los sentidos propuestos como correctos.

4.3.1.3. ESTRUCTURA DEL MÉTODO

Los tres pasos necesarios para poder aplicar el algoritmo presentado en la sección anterior, sobre el recurso ISR-WN, se corresponden con los siguientes:

1. Creación del grafo inicial (o grafo de desambiguación).
2. Particionamiento *N-Cliques*.
3. Selección de sentidos correctos.

Para entender mejor el funcionamiento de la propuesta se utiliza un ejemplo según se muestra en la Figura 31. El proceso comienza a partir de la introducción de una frase inicial, seguido se extraen los lemas de las palabras, y es a partir de aquí cuando comienza el primer paso. A continuación se describen los tres pasos con detalle.

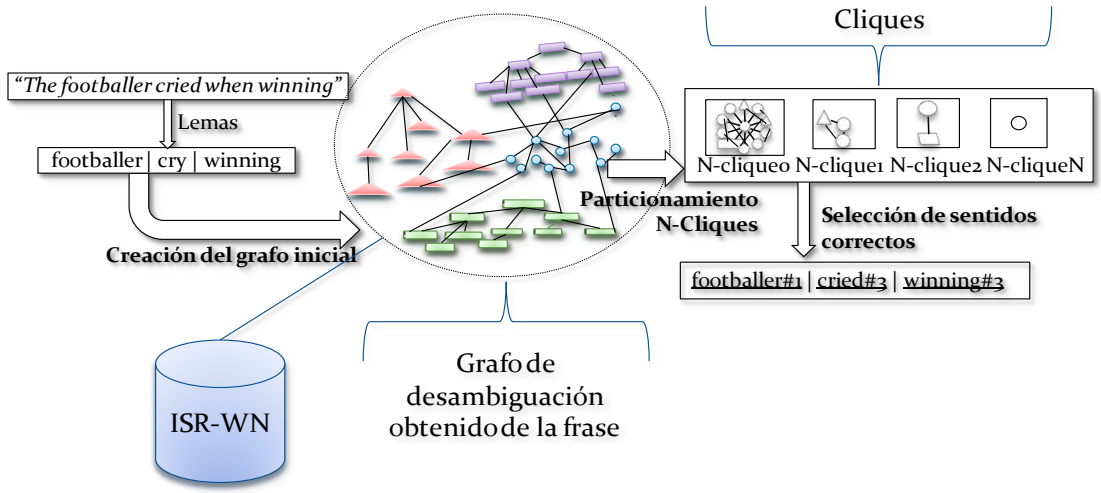


Figura 31. Modelo general de WSD mediante *N-Cliques*.

4.3.1.4. CREACIÓN DEL GRAFO INICIAL

El objetivo de este primer paso es construir un grafo semántico de todos los sentidos de las palabras de cada oración. Para cada texto a analizar se genera un grafo con el fin de descartar elementos que estén fuera del alcance del texto analizado. Este proceso se podría aplicar sobre todo el grafo de ISR-WN, pero sería demasiado costoso. Siguiendo las tendencias de investigadores como (Agirre and Soroa, 2009, Tsatsaronis et al., 2007, Sinha and Mihalcea, 2007, Navigli and Lapata, 2007), se hace más conveniente la creación de pequeños sub-grafos acordes a cada texto a analizar. Para la creación del sub-grafo se toman las conexiones entre sentidos, utilizando la búsqueda del menor camino utilizando BFS (*Breath First Search*) para hallar el camino más corto entre todos los sentidos frente a todos los conceptos relevantes seleccionados.

La base de conocimiento está formada por conceptos (ej. del ISR-WN) y sus relaciones, además del diccionario de WN. El análisis se realiza sobre un grafo completo no dirigido $G = (V, E)$ donde los nodos se encuentran representados por conceptos de la LKB (v_i) y cada relación entre v_i y v_j se representa como una arista $e_{i,j}$. En este caso, se tienen en cuenta las dimensiones de WN, WND, WNA y SUMO para la creación de la red a partir de la información proporcionada por el recurso ISR-WN.

El sub-grafo a generar contiene informaciones representativas de la frase en análisis. Este sub-grafo se conoce como grafo de desambiguación G_D y se construye como se describe a continuación. Si cada k-ésima palabra w^k que pertenece al contexto de la frase L, tiene asociado el conjunto de sentidos sw^k entonces $sw^k = \{sw_1^k, sw_2^k, \dots sw_t^k\}$, tal que $sw^k \subseteq V$ y cada sentido sw^k_j de w^k se identifica como un concepto $v^k_j \in C^k$ tal que $C^k \subseteq V$. C^k se refiere a los conceptos de la palabra w^k hallados en la LKB.

Para deducir el número de sentidos de las palabras que se toman como entrada en la construcción del grafo G_D, se aplica a cada palabra w^k un proceso de categorización gramatical mediante el uso del *Pos-Tagger Freeling*. Este proceso obtiene de L cada palabra categorizada w_p^k, siendo p una categoría gramatical ($p \in POS$), lo que es igual a $w_p^k \in L \times POS$. Si se tiene definido w_p^k, entonces se puede resumir la función de obtención de sentidos de una palabra categorizada $Senses_D(w^k, p)$ como $Senses_D(w_p^k)$. Con el uso de esta función resumida se obtiene entonces el conjunto de sentidos de una palabra categorizada gramaticalmente como el subconjunto de sentidos $sw_p^k = \{sw_p^k{}_1, sw_p^k{}_2, \dots, sw_p^k{}_n\}$, tal que $sw_p^k \subseteq V$.

Se asume además que C' representa el conjunto de conceptos relevantes de la frase L obtenidos por la función $Conceptos\ Relevantes(L)$ tal que $C' \subseteq V$. Esta función varía dependiendo del algoritmo utilizado, en este caso se pueden utilizar dos opciones, RST y *Reuters Vector*. De esta manera, para la construcción de G_D se aplica la obtención del menor camino entre todos los conceptos relevantes C' y los sentidos sw_p^k obtenidos de las palabras categorizadas de L, conociendo que todos estos vértices se encuentran contenidos en el grafo de la LKB (G_{KB}). Se tiene entonces $min_{C'}$ como el conjunto de caminos mínimos de C' respecto a cada vértice v_i tal que $v_i \in V$.

El algoritmo utilizado para hallar el menor camino es BFS y se repite para cada concepto relevante de C' del contexto analizado. Entonces, si $min_{C'}$ es capaz de almacenar los caminos mínimos de todos los conceptos C' obtenidos en L y hacia los todos vértices de G, atravesando múltiples conceptos y relaciones, los vértices y las relaciones que se incluyen en G_D se rigen por el siguiente formalismo, $G_D = \{ \bigcup_{k=1}^{m} min_{C'_i} / C'_i \in sw_p^k \}$. Entonces, este G_D constituye un sub-grafo de G_{KB}.

La construcción de la red semántica inicial de cada frase se describe en la Figura 32. Al extraer los conceptos relevantes se garantiza que los nodos de la red se establezcan alrededor de esta información y se encuentre más centralizada.

Estos son los pasos para construir al grafo inicial o G_D:

(I) Discriminación gramatical teniendo en cuenta la sugerencia del *Pos-Tagger Freeling*.
(II) Obtención de los conceptos relevantes (se aplican dos variantes)
- Aplicación del método *Reuters Vector* (Magnini *et al.*, 2002)
- Aplicación de RST

(III) Obtención de caminos mínimos entre todos los sentidos y los diez conceptos más relevantes. Finalmente, se crea el grafo inicial sin elementos repetidos.

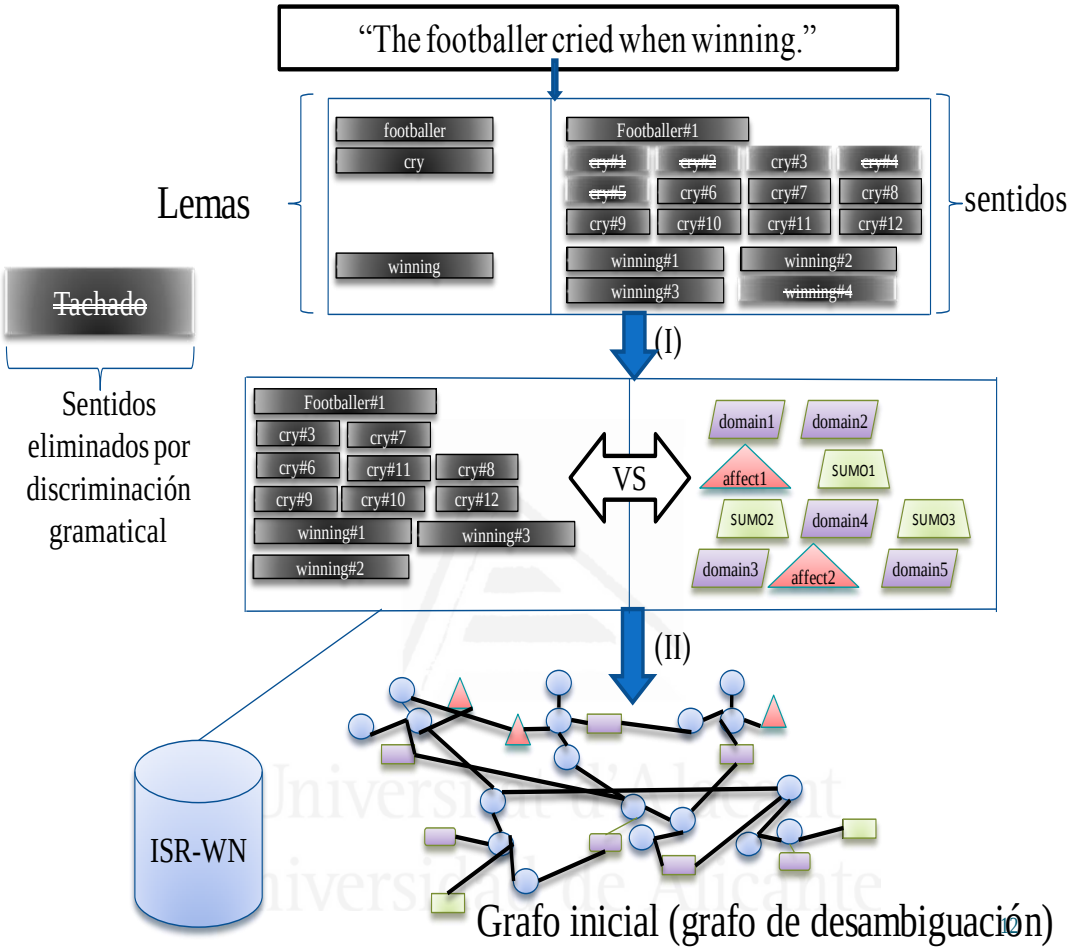


Figura 32. Creación del grafo de desambiguación.

Como resultado se obtiene el G_D compuesto por *synsets* de WN, etiquetas de dominios de WND y WNA, y categorías de SUMO. De la selección de conceptos relevantes se elimina el dominio *Factotum*, debido a que este constituye un dominio genérico, asociando palabras que aparecen en contextos generales, sin ofrecer información relevante (Magnini and Cavaglia, 2000). Además, detrás la evaluación de los resultados de varios experimentos se confirmó que introducía ruido en la clasificación semántica.

A continuación se explica cada una de las opciones evaluadas para obtener los conceptos relevantes. Nótese que para aplicar el método completo, solamente se ha de escoger una de las dos opciones. En la sección de evaluación se muestran experimentaciones con cada una de las opciones a fin de demostrar cuál resulta más efectiva.

4.3.1.4.1. *N-CLIQUES* COMBINADO CON *REUTERS VECTOR* (*N-CLIQUES*+RV)

El método de *Reuters Vector* descrito en la sección 2.5.4.2.1 consiste en un método de adquisición para identificar conceptos o etiquetas semánticas que son relevantes en un *Reuters corpus* [72] para un *synset*. Siguiendo las palabras de Magnini "Como primer paso se construye una lista de lemas relevantes como unión de sinónimos y glosas del *synset.* Esta lista representa el contexto del *synset* en WordNet, y se usará para estimar la probabilidad de un dominio obteniendo un *synset* en el corpus. Esta información se colecciona en un vector llamado *Reuter Vector*, con una dimensión en cada dominio" (Magnini *et al.*, 2002). La única diferencia entre la propuesta original y la aplicada en esta sección, radica en que no se utiliza la información de las glosas, solamente la lista de sinónimos.

Si teniendo una frase f se extrae la lista de palabras $L = \{w^1, w^2, \dots, w^l\}$, las cuales como entrada en el diccionario de WN tienen sus representaciones en la LKB. Esto se traduce en WN $\subseteq$ LKB (por ejemplo tomando como LKB al recurso ISR-WN). De esta manera, a cada palabra w^k se le asocia el conjunto de sentidos $sw^k = \{sw_1^k, sw_2^k, \dots, sw_n^k\}$, donde $sw^k{}_j$ ($1 \leq j \leq n$) representa el j-ésimo sentido de la palabra w^k. Se define entonces la LKB como un grafo completo no dirigido $G = (V, E)$. En este grafo los vértices pueden estar representados por conceptos $C = \{c_1, c_2, \dots, c_r\}$ o por sentidos $S = \{s_1, s_2, \dots, s_m\}$, o sea $V = C \cup S$. La relación entre dos vértices v_i,v_j se representan como una arista $e_{i,j}$. Entonces, a partir de una lista L de palabras es posible obtener un conjunto de elementos semánticos $CS' = C' \cup sw$ siendo C' conceptos conformados por aquellos vértices $v_i \in C$, para los cuales existen aristas $e_{i,j}$ donde $v_j \in sw^k$, tal que los sentidos sw^k se corresponden con las palabras incluidas en L, siendo $sw^k \subseteq S$.

Para cada concepto incluido en la LKB independientemente del tipo de clasificación (dominio. emoción o categoría), es posible calcular las probabilidades de fortaleza relacional que presenten con respecto a cada frase analizada. Para ello, se utiliza la ecuación (43), que a su vez está basada en las ecuaciones (18) y (19) descritas en la sección de Medidas de Similitud Semánticas en concreto *Reuters Vector* (véase la sección 2.5.4.2.1).

$$P(D|f) = \sum_{k=1} \sum_{j=1} P(D|sw_j^k) \tag{43}$$

Donde D es el concepto obtenido (o dominio, correspondiente con $v_j \in CS'$)[73] y f (frase) se corresponde con L. $P(D|f)$ es la probabilidad de distribución conjunta de D respecto a f, k es el índice de la k-ésima palabra, $sw^k{}_j$: es el j-ésimo sentido de la k-ésima palabra y $P(D|sw^k{}_j)$ es la probabilidad de distribución conjunta del concepto D asociado al sentido $sw^k{}_j$, esta probabilidad se calcula con la ecuación (18) y esta a su vez aplica la ecuación (19).

Para obtener la lista de lemas primero se obtienen todos los lemas de los *synsets* de cada palabra de la frase y luego se obtienen las listas de sinónimos asociados a cada uno de los *synsets* (ej. WN1.6 (*entity%1:03:00::* (Lista de lemas (*entity*, *something*), glosa {*anything having existence* (*living or nonliving*)))). La probabilidad $P(D)$ de la ecuación (18) asume que tener valor uno (Magnini *et al.*, 2002).

[72] http://about.reuters.com/researchan/dstandards/corpus/

[73] Donde D es posible valorarlo como concepto o *synset* de WN según el interés del ejecutor de la ecuación.

Originalmente *Reuters Vector* se ha utilizado para medir asociaciones entre *synsets* anotados en corpus de dominios específicos. En esta ocasión se aplica para *synsets* integrados relacionalmente en el recurso ISR-WN. Como resultado se obtiene un conjunto de conceptos relevantes CR que conceptualizan la frase, correspondiéndose CR con el C' requerido la sección 4.3.1.4 de creación del grafo inicial de desambiguación.

4.3.1.4.2. *N-CLIQUES* COMBINADO CON ÁRBOLES SEMÁNTICOS RELEVANTES (*N-CLIQUES*+RST)

En esta ocasión, se aplican los árboles semánticos relevantes RST's explicados en la sección 4.2.2.2.1. Es importante resaltar que solamente se obtienen los árboles semánticos, los demás pasos de desambiguación de la propuesta de RST se omiten. Como resultado, todos los conceptos relevantes de todos los recursos quedarán agrupados en una colección correspondiéndose con el CS' requerido la sección 4.3.1.4 de creación del grafo inicial de desambiguación.

4.3.1.5. PARTICIONAMIENTO *N-CLIQUES*

Una vez obtenido el grafo inicial se utiliza como entrada para el algoritmo de Particionamiento de *N-Cliques*. Para ello, se obtienen varios conjuntos de nodos representados por los diferentes tipos de elementos que componen el grafo inicial.

4.3.1.6. SELECCIÓN DE SENTIDOS CORRECTOS

Para este paso, primero se ordenan los conjuntos particionados en *N-Cliques* por la cantidad de vértices que los constituyen. Luego, para cada correspondiente lema de la frase en cuestión, se buscan los *synsets* que lo representan en cada *Clique*. Si no existe ningún *synset* en el primer *N-Clique* se pasa al siguiente, y así sucesivamente. El *synset* hallado en este proceso se considera como correcto. Pero en caso de encontrar dos o más *synsets* en un mismo *N-Clique* asociados al mismo lema, se selecciona el *synset* de mayor frecuencia de aparición según sugiere la herramienta *Freeling* (Atserias *et al.*, 2006). Se tiene en cuenta una excepción con respecto al lema *be*, para este lema siempre se selecciona como sentido correcto el más frecuente (ej. *be%2:42:03::*). Esto se debe a que este verbo es uno de los más polisémicos y además no se inclina a ser utilizado en contextos específicos.

4.3.2. *PAGERANK* COMBINADO CON FRECUENCIAS DE SENTIDOS

Como se ha mencionado anteriormente el célebre algoritmo *PageRank* introducido por (Brin and Page, 1998), es un método para la clasificación de vértices dentro de un grafo de acuerdo a su importancia estructural. La idea principal de *PageRank* es que cada vez que un enlace de v_i a v_j existe en un grafo, se produce un voto del nodo i al nodo j, y como consecuencia aumenta la importancia del nodo j. Además, la fuerza de los votos de i a j también depende de la importancia que presente el nodo votante i. La filosofía que se persigue, es que el nodo más importante es el que más fuerza tenga en sus votos. Por otra parte, el modo de ejecutar el *PageRank* se percibe como el resultado de un proceso de paseo aleatorio dentro de la red, donde la importancia final del nodo i representa la probabilidad de paseo al azar sobre el grafo, terminando cada ciclo en el nodo i.

La representación general se plantea de la siguiente forma. Sea G un grafo con N vértices $v_1, \ldots, v_N$ y d_i el grado de salida (número de relaciones) del nodo i. Entonces el cálculo del vector de *PageRank* (**Pr)** aplicado sobre el grafo G se resuelve con la siguiente ecuación tomada

de (Agirre and Soroa, 2009) y adaptada al contexto de WSD, descartando además los nodos que no presentan aristas.

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)v \quad (44)$$

Donde:

- M es una matriz de transición probabilística $M \times N$, donde $M_{j,i} = \frac{1}{d_i}$ si existe un vínculo entre el nodo i y el j, y de no ser, sería igual a cero.
- v es un vector $N \times 1$ cuyos elementos equivalen al valor resultante de $\frac{1}{N}$.
- c es conocido como factor de amortiguación (*damping factor*), que comprende un valor escalar entre el rango $\{0 \ ... \ 1\}$. Comúnmente se eligen valores entre 0.85 y 0.95.
- **Pr** representa, en términos generales, el vector de probabilidades obtenidas en el paseo al azar de llegar a cualquier nodo.

En cada paso de la iteración se combinan c, M y **Pr**. El término $(1-c)v$, puede ser visto como un factor de suavizado que garantiza que converja el cálculo de *PageRank* de una distribución estacionaria única (Agirre and Soroa, 2009). Como valor inicial del vector de pesos **Pr** se le asignan los valores resultantes de $\frac{1}{N}$, con el fin de iniciar con rangos igualitarios para todos los nodos.

Se debe prestar importante atención a los valores con los que se inicia v. Tradicionalmente se les asigna $\frac{1}{N}$, sin embargo analizando el punto de vista de (Haveliwala, 2002, Agirre and Soroa, 2009) la evaluación del vector v admite asociar mayor importancia a algunos vértices respecto a otros (la importancia será decidida según el interés del ejecutor de *PageRank*). Siguiendo la filosofía de *PageRank* donde el nodo más importante es el que más fuerza tenga en sus votos, entonces, la fuerza inicial de los votos de un nodo i hacia un nodo j se verá favorecida por la importancia que presente el nodo votante i.

Motivados por esta idea de ponderar algunos nodos del vector v, varios autores han aplicado el algoritmo de *PageRank* en WSD. Por ejemplo, (Sinha and Mihalcea, 2007) luego de calcular seis diferentes medidas de similitud entre elementos de la red, evalúa los valores obtenidos en v; (Agirre and Soroa, 2009) se inclinan en esa dirección, proponiendo *Personalizing PageRank* mediante la concentración del peso de probabilidad inicial de manera uniforme sobre los nodos de la palabra de reciente introducción (sobre los *synsets* asociados a las palabras del texto analizado). Entonces, el vector resultante puede ser visto como una medida de la relevancia de los conceptos estructurales en correspondencia al contexto de entrada. En esta sección se propone hacer uso de *Personalizing PageRank* ponderando solamente los sentidos correspondientes a la frase analizada en el vector *v* haciendo uso de los valores normalizados de frecuencias de sentidos. Además se toma como bases de conocimiento a todos los recursos presentes en ISR-WN (con excepción de SWN), conjuntamente con los utilizados por (Agirre and Soroa, 2009), en concreto *eXtended WordNet* 1.7 y 3.0, y las relaciones entre pares de sentidos que coocurren en el corpus de SemCor.

4.3.2.1. PROPUESTA DE WSD APLICANDO *PERSONALIZING PAGERANK* COMBINADO CON FRECUENCIAS DE SENTIDOS

En esta sección se presenta la aplicación de Ppr+Frec (*Personalizing PageRank* combinado con Frecuencias de sentidos). La propuesta descarta el uso tradicional del algoritmo, que consiste en aplicarlo sin tener en cuenta la distinción de relevancia inicial entre sentidos. Además, siguiendo las nuevas tendencias de aproximaciones de WSD basadas en grafos, se genera para cada frase de entrada un grafo de desambiguación. En este marco, es necesario para

clasificar a los sentidos de las palabras de acuerdo con las otras palabras que conforman el contexto. El método general se ilustra en la Figura 33 y se divide en los siguientes pasos:

1. Creación del grafo inicial.
2. Aplicación de Ppr+Frec sobre el sub-grafo generado.
3. Selección de sentidos correctos.

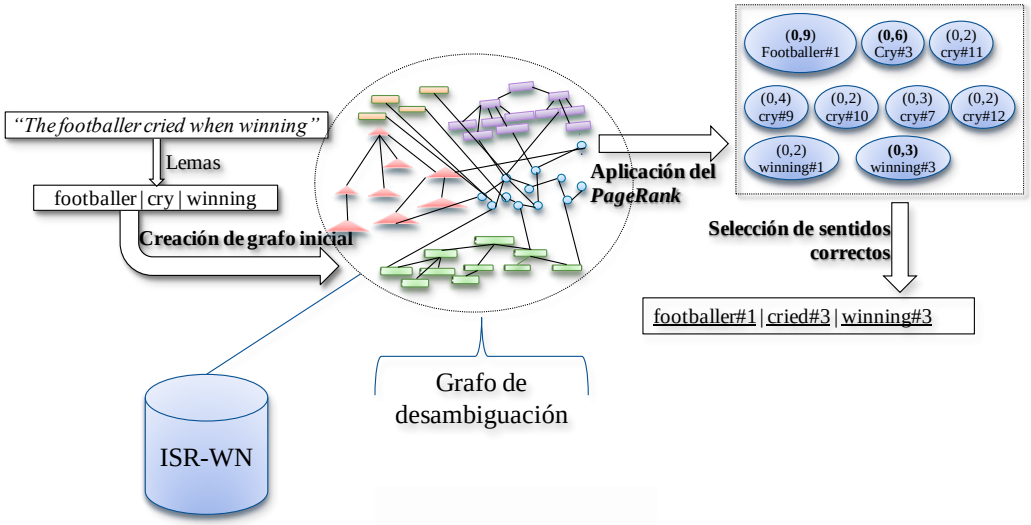


Figura 33. Modelo general de WSD mediante Ppr+Frec.

Según se muestra en la Figura 33 todo comienza cuando una frase es introducida. Luego, se obtienen los lemas de las palabras continuando con tres pasos siguientes que a continuación se describen en detalle.

4.3.2.1.1. CREACIÓN DEL GRAFO INICIAL

El objetivo de este paso consiste en construir un grafo semántico de todos los sentidos de las palabras en cada oración. Para ello se genera un sub-grafo G_D para cada texto a analizar según se ha comentado en secciones anteriores. Para la creación del sub-grafo se toman las conexiones entre elementos de la base de conocimiento utilizando el camino más corto, ahora entre todos los sentidos. La construcción de la red semántica inicial de cada frase se describe en la Figura 34. La LKB se conforma por conceptos y sus relaciones además del diccionario de WN. Se trabaja entonces, sobre un grafo completo no dirigido $G = (V, E)$ donde los nodos se encuentran representados por conceptos de la LKB (v_i) y cada relación entre v_i y v_j se representa como $e_{i,j}$. Entre las redes de conocimiento que se tienen en cuenta están:

- ISR-WN enriquecido (se añade la dimensión de SC).
- eXtended WordNet 1.7 (XWN1.7).
- glosas desambiguadas de WN 3.0 (XWN3.0).
- relaciones entre pares de sentidos que coocurren en el corpus de SemCor (pSemcor).

Al utilizar como base de conocimiento WN +XWN1.7, para poder distinguirla ahora se nombra LKB1.7 y en caso de ser WN +XWN3.0 es LKB3.0.

A partir de una frase de entrada se extrae la lista de palabras $L = \{w^1, w^2, \ldots, w^l\}$, contenido, las cuales como entrada en el diccionario de WN tienen sus representaciones en la LKB además de sus relaciones con conceptos según se ha explicado en secciones anteriores, se tienen los conceptos $C^k = \{v^k{}_1, v^k{}_2, \ldots, v^k{}_m\}$ asociados a la palabra w^k en la red LKB. Como resultado del proceso de desambiguación cada concepto de la LKB (incluyendo los nodos que representan sentidos) tiene asociado un valor de importancia. Entonces, para cada palabra a desambiguar asociada al grafo G, solamente se tiene que elegir el sentido asociado a ella que maximice su

valor. La frase de entrada en esta propuesta no se ha restringido de acuerdo a la ventana de palabras, haciendo posible introducir todo texto que constituya una frase.

Estos son los pasos para construir al grafo inicial:

(I) Discriminación gramatical teniendo en cuenta la sugerencia del corpus de Senseval convertido a formato SemCor (disponible el sitio de Rada Mihalcea[74])

(II) Obtención de caminos mínimos entre todos los sentidos y luego se crea el grafo inicial sin elementos repetidos.

El sub-grafo a generar contiene informaciones representativas de la frase en análisis. Este sub-grafo se conoce como grafo de desambiguación G_D, que se define como $G_D \subseteq G$ y se construye como se describe a continuación. A cada palabra w^k que pertenece al contexto de la frase se comprende por tener asociados sentidos $sw^k = \{sw^k_1, sw^k_2, \ldots sw^k_p\}$.

Es importante señalar que la LKB a utilizar se comprende por WN, etiquetas de dominios de WND, WNA y SC, y categorías de SUMO. Además de las relaciones proporcionadas por XWN1.7, XWN3.0 y pSemcor. Con la intervención de todos estos recursos, se hace más compleja la representación de la LKB.

Con el fin de cumplir el paso (I) del proceso de WSD, una palabra w^k se categoriza gramaticalmente según las sugerencias del corpus de Senseval, obteniéndose w_p^k, siendo p una categoría gramatical ($p \in POS$), lo que es igual a $w_p^k \in L \times POS$. Si se tiene definido w_p^k, entonces se puede resumir la función $Senses_D(w^k, p)$ como $Senses_D(w_p^k)$, obteniendo el conjunto de sentidos de una palabra categorizada gramaticalmente como el subconjunto $sw_p^k = \{sw_{p\,1}^k, sw_{p\,2}^k, \ldots, sw_{p\,n}^k\}$. Para obtener el conjunto de sentidos categorizados de todas las palabras de L, tal que $Lp = L \times POS$ se aplica la función $Senses_D(Lp)$. Esta función posibilita que para la construcción del grafo G_D se utilice la obtención del menor camino entre todos los vértices $v_p^k \in Senses_D(Lp)$, siendo cada sentido $sw_{p\,j}^k$ un concepto $v_p^k \in C_p^k$ (Conceptos asociados w_p^k). Nótese que el conjunto de conceptos C_p^k está contenido en el grafo de la LKB (G_{KB}) donde cada vértice $v_p^h \in \cup_{h \neq k} C_p^h$.

Se tiene entonces con relación al paso (II) que $min_{v_p^k}$ es el conjunto de caminos mínimos del vértice v_p^k hacia todos los elementos del grafo G_{KB}. Si $min_{v_p^k}$ es capaz de almacenar los caminos mínimos entre todos los sentidos de la palabra categorizada w_p^k, atravesando múltiples conceptos y relaciones, los vértices y las relaciones que se incluyen en el grafo G_D se rigen por el siguiente formalismo, $G_D = \{ \cup_{k=1}^{l} min_{v_p^h} / v_p^h \in C_p^k \}$. Entonces, el grafo G_D constituye un sub-grafo de G_{KB}. El algoritmo utilizado para hallar el menor camino es *BFS* (*Breath First Search*) y se repite para cada sentido de cada palabra del contexto analizado.

Como resultado se obtiene el grafo inicial compuesto por *synsets* de WN, etiquetas de dominios de WND, WNA y SC, y categorías de SUMO, además de las relaciones proporcionadas por XWN1.7, XWN3.0 y pSemcor. Nótese que es posible al momento de realizar la evaluación poder seleccionar qué recursos forman parte de la LKB. Es importante resaltar que en caso de seleccionar WND dentro de la LKB se elimina el dominio *Factotum*, esto sucede debido a que este constituye un dominio genérico asociando palabra que aparecen en contextos generales, no ofreciendo información relevante (Magnini and Cavaglia, 2000). Y al aplicar caminos mínimos reduce mucho las distancias entre elementos de G_{KB} sin proporcionar semántica relevante.

[74] http://www.cse.unt.edu/~rada/downloads.html

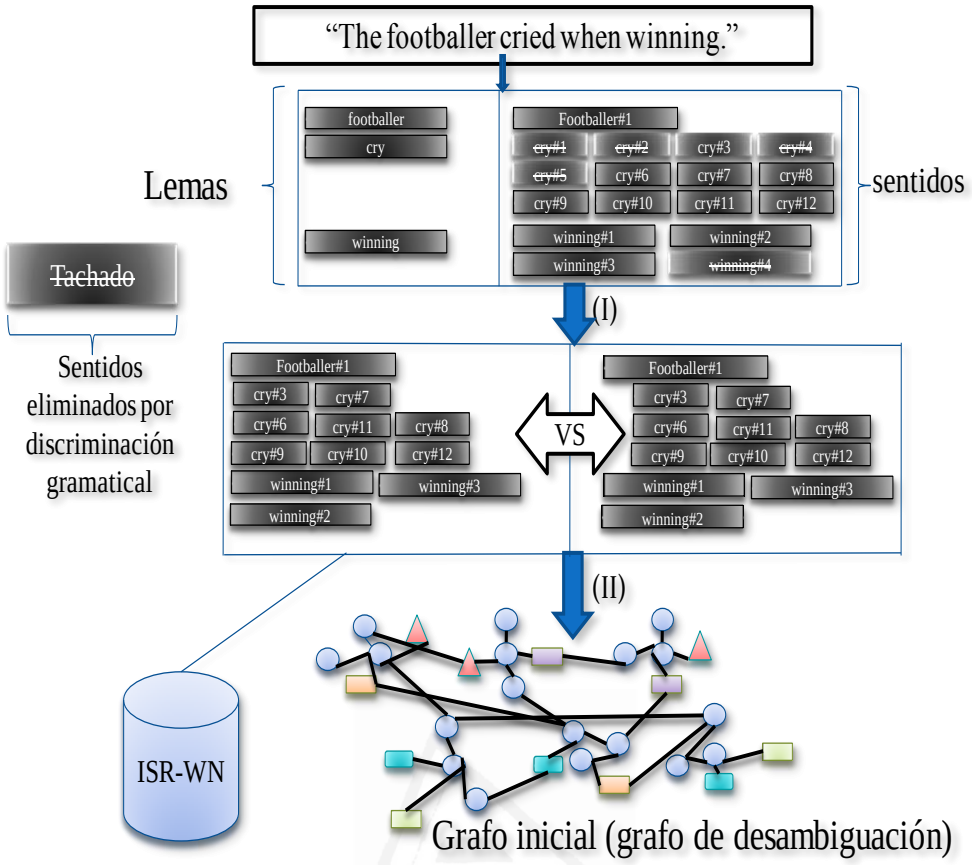


Figura 34. Creación del grafo de desambiguación.

4.3.2.1.2. APLICACIÓN DE PPR+FREC SOBRE EL SUB-GRAFO GENERADO.

Una vez que es construido G_D se aplica el algoritmo de Ppr+Frec sobre este. Para ello se introducen cada una de las frases sin importar las ventanas de palabras que estas constituyan. En esta ocasión se propone hacer uso de *PageRank* ponderando únicamente los sentidos correspondientes a la frase analizada en vector v. Para ello, se utilizan los valores normalizados de **frecuencias de sentidos**. Entonces, todos los vértices en G_D se inicializan con valor $\frac{1}{N}$ a excepción de los vértices que representan sentidos de las palabras de la frase analizada, estos tienen valores del rango {0 ... 1} asociados a su frecuencia de aparición en SemCor.

En esta propuesta y siguiendo las sugerencias de (Agirre and Soroa, 2009) se establece un límite de 30 iteraciones, pues se ha comprobado que a partir de ese ciclo los valores adquiridos por **Pr** mantienen muy poca variación. Como valor constante de *damping factor* se asigna 0.85.

4.3.2.1.3. SELECCIÓN DE SENTIDOS CORRECTOS

Tras el proceso anterior se obtiene un vector representativo de cada sentido a partir de los vértices pertenecientes a G_D. Entre ellos se encuentran los sentidos particulares de cada palabra del contexto. Para la selección de los sentidos correctos, basta con elegir de cada palabra objetivo el sentido correspondiente que maximice su valor de rango. Aquí se descartan los empates, debido a que durante la experimentación no existió ninguna aparición de este fenómeno, tras haber ponderado de vector v con frecuencias normalizadas de sentidos. Así también, se descarta que existan al menos dos sentidos que asociados a los mismos conceptos obtengan el mismo valor de rango.

4.4. EVALUACIONES Y RESULTADOS

En esta sección se muestran las evaluaciones correspondientes a las propuestas de WSD descritas en este capítulo. Con el objetivo de establecer comparaciones y conocer en qué medida se han superado o disminuido los resultados alcanzados, se tienen en consideración descripciones, reportes emitidos por otros autores y competiciones de la rama. A continuación, se detallan resultados referentes a cada propuesta del capítulo, luego una comparativa entre sistemas considerados relevantes y por último criterios notables contrastantes con competiciones de WSD.

4.4.1. ÁRBOLES SEMÁNTICOS RELEVANTES

El proceso de experimentación del método basado en la obtención de Árboles Semánticos Relevantes, se aplica sobre en el corpus de Senseval-2 para la tarea *English All Words* y *English All Words on Specific Domain* de Semeval-2. Seguidamente en las dos secciones siguientes se analizan los correspondientes resultados y comportamientos.

4.4.1.1. EVALUACIÓN SOBRE EL CORPUS DE SENSEVAL-2

Sobre el corpus de Senseval-2 se realizan varias pruebas con el objetivo de identificar elementos básicos que marcan el curso de la experimentación. Inicialmente se efectuaron experimentos de WSD sin utilizar la discriminación gramatical que proporciona *Freeling*, obteniendo un valor de *Recall* de un 34%. El modo de seleccionar el sentido correcto también ha estado sujeto a experimentos, para ello se utilizan las medidas del coseno del ángulo entre vectores (a continuación en la ecuación (45)), matriz de coocurrencia, sumatorio de conceptos relevantes y un método de votación básico que los abarca a los tres. De estos, las mejores aproximaciones estuvieron dadas por el sumatorio de *AR* descrito en la sección 4.2.1.3. A continuación solamente se ilustran resultados que aportan información suficiente como para dilucidar el camino a seguir y mejorar los resultados de WSD, el resto (matriz de coocurrencia y método de votación) no se abordan.

Por ejemplo, al aplicar coseno con la ecuación (4) y adaptarla al contexto en cuestión, quedaría tal y como la presenta la ecuación para medir similitud entre los vectores del contexto y de los sentidos. Los resultados no sobrepasan el 32% de *Recall*.

$$cos(\overrightarrow{Vs}, \overrightarrow{Vf}) = \frac{\sum_{i=1}(Vs_i * Vf_i)}{\sqrt{\sum_{i=1} Vs_i^2} * \sqrt{\sum_{i=1} Vf_i^2}}) \quad (45)$$

Donde:

- $\overrightarrow{Vs}$: es un vector *AR* del sentido.
- $\overrightarrow{Vf}$: es el un vector *AR* de la frase.
- Vs_i: accede al valor i-ésimo de *AR* de Vs
- Vf_i: accede al valor i-ésimo de *AR* de Vf

El acumulado de similaridad total entre cada sentido y la frase al utilizar los recursos WN, WND, WNA y SUMO queda de la siguiente forma:

$$ACTs = \frac{\cos(\overrightarrow{Vss}, \overrightarrow{VARs}) + \cos(\overrightarrow{Vds}, \overrightarrow{VARd}) + \cos(\overrightarrow{Vas}, \overrightarrow{VARa}) + \cos(\overrightarrow{Vwns}, \overrightarrow{VARwn})}{4} \quad (46)$$

Para poder aplicar la fórmula del coseno, los vectores Vss,Vds, Vas y $Vwns$ asumen su formato de la siguiente forma [valor AR| concepto].

A medida que avanza el proceso de experimentación surgen varios problemas referentes a *Freeling*. Estos problemas han sido mencionados con anterioridad en la sección 4.2.1.1, pero sería correcto hacer la salvedad, de que este método de WSD es dependiente de cierta forma de un analizador gramatical.

Al aplicar las pruebas de la propuesta descrita en la sección 4.2.1, sobre todo el corpus de Senseval-2 para la tarea *English All Words*, se tienen en consideración la inclusión individual y combinada de los recursos WN, WND, WNA y SUMO. Los resultados se muestran en la Tabla 23.

Tabla 23. Evaluación de RST sobre el corpus de Senseval-2.

Experimentos	**Recursos**					***Precision***	***Recall***
	Frecuencias sentidos	WN	SUMO	WND	WNA		
Exp 1 MFS *Freeling*	X					0.412	0.411
Exp 2				X		0.400	0.399
Exp **3**					**X**	**0.424**	**0.423**
Exp 4			X			0.379	0.378
Exp 5		X				0.377	0.376
Exp 6			X	X		0.397	0.397
Exp **7**		**X**	**X**	**X**	**X**	**0.425**	**0.424**

Como se puede apreciar los mejores resultados se obtienen primeramente con el *baseline* solamente analizando por frecuencia. Cuando se decide por frecuencia no se aplica el método descrito, sino que solamente se toma el sentido más frecuente de la palabra objetivo. Las frecuencias de palabras en esta evaluación, se obtienen de la herramienta *Freeling*. Luego, le siguen los reportes que usan WNA. Cabe destacar, que únicamente los *synsets* que denotan emociones se afectan por WNA y una gran mayoría no lo está. Como consecuencia de eso, al no existir una salida del desambiguador, por defecto se asigna el sentido más frecuente. Esto indica que en esta variante está muy presente la influencia de la decisión de MFS. La combinación de los cuatro recursos resulta con el mayor *Recall* de un 42.2% empatada con la inclusión de WNA. Es decir, al aplicar sobre un corpus de dominio general, la combinación de todos los recursos obtiene las mejores respuestas. A continuación se aplica sobre un dominio específico y se verá cuales son los resultados para contrastarlos finalmente.

4.4.1.2. EVALUACIÓN SOBRE EL CORPUS DE SEMEVAL-2

Sobre el corpus de Semeval-2 también se evalúa el método, pero en esta ocasión como sistema participante. El mejor de los resultados mostrados en la Tabla 24 representa el *baseline* MFS, seguido de este está el Experimento 4 con valores bien diferenciados, pero mejores que el resto de los experimentos. Se puede apreciar que en esta evaluación, la integración de recursos no es superior al resto, sucede que este corpus tiene la característica de ser de dominio específico, lo que indica, que el uso de WND se enfoca mejor que el resto de recursos.

Estos resultados obtenidos (considerados bajos por el autor de la Tesis) generan un punto de partida para nuevas investigaciones. Se hace necesario introducir nuevas variantes que hagan válido el uso de la información multidimensional y poder alcanzar resultados superiores. Nótese, que en estas experimentaciones no se tiene en cuenta a la taxonomía de WN, debido a que en el proceso de evaluación el árbol semántico concebido se hace demasiado general. Esto implica que se identifique muy poca información relevante y por consiguiente muy bajos resultados. En la siguiente tabla se muestra el lugar que logra ocupar el sistema que aplica RST evaluado en Semeval-2 (Gutiérrez *et al.*, 2010b).

Experimentos	Recursos				*Precision*	*Recall*
	Frecuencias sentidos	SUMO	WND	WNA		
Exp 1(MFS Semeval-2)	X				0.505	0.505
Exp 2				X	0.242	0.237
Exp 3		X			0.267	0.261
Exp 4			**X**		**0.328**	**0.322**
Exp 5		X	X		0.308	0.301
Exp 6		X	X	X	0.308	0.301

Tabla 24. Evaluación de RST en Semeval-2.

El comportamiento de la clasificación de cada categoría gramatical para la mejor de las propuesta presentadas en la Tabla 24, se muestra en la Tabla 25. Estos resultados todavía están muy lejos de poder ser fiables, pero sí identifican que el análisis semántico con árboles relevantes es válido en WSD, solamente se necesita introducirle otro tipo de información con tal de mejorar la exactitud.

Categoría gramatical	*Precision*	*Recall*
sustantivo	0.335	0.335
verbo	0.284	0.284

Tabla 25. Resultados de las categorías gramaticales en Semeval-2 de RST.

Rank	*Precision*	*Recall*	*Type*	*Rank*	*Precision*	*Recall*	*Type*
1	0.570	0.555	WS	…	….	…	
2	0.554	0.540	WS	…	…	…	
3	0.534	0.528	WS	26	0.370	0.345	WS
4	0.522	0.516	WS	**27 Exp 4(RST)**	**0.328**	**0.322**	**KB**
5	0.513	0.513	S	28	0.321	0.315	KB
MFS	0.505	0.505	-	29	0.312	0.303	KB
…	…	…		*Random*	0.23	0.23	

Tabla 26. *Ranking* de Semeval-2 con evaluaciones de RST (débilmente supervisado (WS), supervisado (S), basado en conocimiento (KB)).

4.4.1.3. ANÁLISIS GENERAL DEL PROCESO DE WSD AL USAR ÁRBOLES SEMÁNTICOS RELEVANTES

Luego de haber analizado el comportamiento de la propuesta de RST sobre los corpus de Senseval-2 y Semeval-2, se obtienen ciertas conclusiones válidas para continuar la investigación y lograr mejores aproximaciones. Esta propuesta es capaz de obtener información conceptual relevante de los textos y representarla de forma jerárquica. Dicha característica posibilita obtener una visión más o menos abstracta del texto según el interés del ejecutor. Quizás en su utilización en WSD, sea necesario vincular esa abstracción contextual, con otro tipo de información capaz de precisar con mejor exactitud la clasificación de palabras. Con RST hasta ahora se consigue colocar un texto en un contexto determinado por mediación de conceptos. ¿Pero cuántos sentidos para cada palabra no pueden coexistir conjuntamente en contextos similares? Para responder a esa pregunta se toma como ejemplo la Tabla 4, donde se representa la palabra del inglés *man* asociada a varios dominios de WND. Como se puede apreciar el uso de etiquetas de dominios o conceptos en general consiguen la discriminación de ciertos sentidos de la palabra, pero incluso así continua siendo polisémica en muchos casos. Ahora el reto, es lograr discernir del sub-conjunto se sentidos asociados a los RST los cuales adquieren mayor importancia entre sí. Uno de los detalles importantes revelados en estas evaluaciones, es que para los casos que se desea desambiguar texto de dominio específico, el recurso WND aislado, funciona con mejor exactitud que los demás. Por otra parte, si se desea analizar texto de dominio general, la combinación de dimensiones es la mejor opción.

4.4.2. ÁRBOLES SEMÁNTICOS RELEVANTES COMBINADOS CON FRECUENCIAS DE SENTIDOS

En esta sección se desea confirmar si la combinación conceptual (RST) y estadística de frecuencias de sentidos, supera los resultados de MFS y RST por individual. Esto resulta en un método de WSD que busca conocer cuáles son los sentidos más frecuentes según un contexto determinado. Para ello, se ejecutan evaluaciones basadas en el corpus de Senseval-2 de la tarea *"English All Words"* y el de Semeval-2 en *"All Words English All Words on Specific Domain"*. El objetivo principal de estas comprobaciones es demostrar cuanto mejoran o empeoran los resultados del RST original y la frecuencia de sentidos si se combinan adecuadamente en un sistema.

4.4.2.1. EVALUACIÓN CON EL CORPUS DE SENSEVAL-2

En primer lugar se analiza en qué medida el incremento de información de frecuencias de sentidos en la variable $Frec_s$, afecta a los resultados en comparación con el RST original. Se debe recordar que esta variable ha sido introducida en la ecuación (41) para calcular los valores acumulados (AC) de cada sentido. Para hacer estas comprobaciones se utiliza únicamente uno de los tres ficheros (d00.txt) del corpus de Senseval-2 y se conducen los siguientes experimentos teniendo en cuenta todos los recursos semánticos.

- Exp 1: Adición a AC un 0% de $Frec_s$ (esto coincide con la propuesta original de RST pero con decisión entre sentidos por votación).
- Exp 2: Adición a AC un 50% de $Frec_s$.
- Exp 3: Adición a AC un 100% de $Frec_s$ (aquí se coloca el 100% de la información de frecuencias de sentidos ya normalizadas).

En la propuesta original de RST de la sección 4.2.1 el valor de AC es calculado para cada dimensión y se suman todos los valores. Esto se hace con el objetivo de obtener un único valor acumulado que combina todos los recursos. En esta propuesta la modificación está en añadir al cálculo anterior de *AC* el valor de $Frec_s$ y aplicar una votación en vez de sumatoria de valores. La Tabla 27 muestra cómo cada evaluación incrementa sus exactitudes a medida que aumenta el grado de información de $Frec_s$. Nótese, que no se mantiene el incremento de pesaje (ej. 150%, 200%, etc) porque la propuesta se convertiría en el proceso de selección de MFS y no eso lo que se desea.

Con el objetivo de determinar en qué medida la inclusión de $Frec_s$ en la propuesta de RST mejora los resultados del *baseline* MFS, se idean nuevos experimentos. A continuación se evalúa la propuesta de adición del 100% de $Frec_s$ a *AC*, pero solamente teniendo en cuenta una dimensión por separado. Esto se hace para conocer qué dimensión es la que más influye en la obtención de los mayores resultados.

- Exp 4: Con el uso del *baseline* MFS con la información de $Frec_s$
- Exp 5: Con el uso de WND
- Exp 6: Con el uso de SUMO
- Exp 7: Con el uso de WNA
- Exp 8: Con el uso de la taxonomía del recurso WN

Luego de realizar estos experimentos se es capaz de identificar cuál dimensión actúa mejor en esta nueva propuesta a favor de la exactitud. Como se puede ver en la Tabla 27, estos cinco experimentos obtienen prometedores resultados.

Otra de las evaluaciones realizadas, es la que combina en un proceso de votación descrito por la ecuación (42) todas las dimensiones (incluyendo MFS). Esta en concreto utiliza los resultados

de estos cinco experimentos como votantes. Entonces el proceso de votación envuelve a los resultados del Exp 4, Exp 5, Exp 6, Exp 7 y Exp 8. Este quedaría descrito de la siguiente forma:

- Exp 9: Proceso de votación entre el Exp 4, Exp 5, Exp 6, Exp 7 y Exp 8.

La Tabla 27 muestra los resultados de estas experimentaciones, donde el Exp 4 representa a la aproximación MFS y además se presentan un conjunto de experimentos que la superan. Además se puede apreciar que el proceso de votación (Exp 9) obtiene los mejores resultados evaluados sobre el fichero d00.txt.

		Votación en Dimensiones (Recursos)					**%Frec$_s$**	***Precision***	***Recall***
Ficheros	Recursos	WN	WNA	SUMO	WND	MFS			
	Experimentos								
d00.txt	Exp 1	X	X	X	X	X	0%	0.408	0.407
	Exp 2	X	X	X	X	X	50%	0.490	0.490
	Exp 3	X	X	X	X	X	100%	0.535	0.534
	Exp 4 (MFS)					X	100%	0.565	0.564
	Exp 5				**X**		**100%**	**0.572**	**0.572**
	Exp 6			X			100%	0.561	0.560
	Exp 7		X				100%	0.555	0.554
	Exp 8	**X**					**100%**	**0.572**	**0.572**
	Exp 9	**X**	**X**	**X**	**X**	**X**	**100%**	**0.575**	**0.575**
Corpus completo (d00.txt, d01.txt, d02.txt)	Exp 10 (MFS)					X	100%	0.601	0.599
	Exp 11	**X**	**X**	**X**	**X**	**X**	**100%**	**0.610**	**0.609**

Tabla 27. Resultados sobre Senseval-2.

Después de identificar la mejor de las combinaciones, se procede a avaluar el proceso de votación y el *baseline* MFS sobre el corpus completo de Senseval-2. Para ello se plantean dos experimentos.

- Exp 10: MFS usando la información de $Frec_s$.
- Exp 11: Votación con las cinco dimensiones (WN, WND, WNA, SUMO, MFS).

En la Tabla 28 se muestra el *ranking* reducido de Senseval-2 para establecer una comparativa entre los resultados obtenidos y los mejores posicionados a nivel mundial. Entre los resultados se aprecia la evaluación de MFS emitida en el artículo de (Preiss, 2006). Estos resultados son diferentes a los obtenidos por el *baseline* que utiliza $Frec_s$. Se debe a que el *baseline* de Preiss utiliza como base de información el fichero *cntlist* de WN 1.7 y el ejecutado en esta propuesta (Exp 10) se basa en *cntlist* de WN 1.6, ambos son extraídos del análisis de SemCor pero a medida que surgen versiones de este los recursos estadísticos se van enriqueciendo.

Rank	***Precision***	***Recall***	***Type***	***Rank***	***Precision***	***Recall***	***Type***
1	0.690	0.690	S	**Exp 11(RST+Frec)**	**0.610**	**0.609**	**U**
MFS	0.669	0.646	-	Exp 10 MFS	0.601	0.599	-
2	0.636	0.636	S	4	0.575	0.569	U
3	0.618	0.618	S	..	..	..	

Tabla 28. *Ranking* de Senseval-2 con evaluaciones de RST+Frec.

Como se puede observar el Exp 11 supera al *baseline* de MFS del Exp 10 sobre todo el corpus analizado. Pero el *baseline* emitido por Preiss es mayor aún. Esto indica que la propuesta RST+Frec logra mejorar el MFS que utilice. Por ese motivo, para obtener mejores resultados se necesitaría experimentar RST+Frec con mejores recursos de frecuencia de sentidos.

4.4.2.2. EVALUACIÓN CON EL CORPUS DE SEMEVAL-2

Esta propuesta se evalúa ahora sobre el corpus de Semeval-2, con el objetivo de comprobar si es capaz de superar el RST original el cual ha formado parte se esta competición. Como se puede observar en la Tabla 29, el proceso de votaciones de todas las dimensiones semánticas logra obtener un 52.7% y 51.5% de *Precision* y *Recall* respectivamente, superando el *baseline* MFS en un 1% y al RST original en un 19.3%.

Rank	*Precision*	*Recall*	*Type*	*Rank*	*Precision*	*Recall*	*Type*
1	0.570	0.555	WS	…	….	…	
2	0.554	0.540	WS	…	…	…	
3	0.534	0.528	WS	26	0.370	0.345	WS
4	0.522	0.516	WS	**27** Exp 4 **(RST)**	**0.328**	**0.322**	**KB**
(RST+Frec)	**0,527**	**0,515**	**KB**	28	0.321	0.315	KB
5	0.513	0.513	S	29	0.312	0.303	KB
MFS	0.505	0.505	-	*Random*	0.23	0.23	

Tabla 29. *Ranking* de Semeval-2 con evaluaciones de RST+ Frec.

En esta competición únicamente se evalúan los sustantivos y los verbos. Según los resultados de la Tabla 30, en cada categoría gramatical se establecen aciertos cercanos a los mejores resultados logrados en dicha competición y publicados en (Agirre *et al.*, 2010).

Categoría gramatical	*Precision*	*Recall*
sustantivo	0.544	0.537
verbo	0.494	0.454

Tabla 30. Resultados de las categorías gramaticales en Semeval-2 de RST+Frec.

Con relación a determinar cuánto ruido pudiera introducir la herramienta *Freeling* como *Pos-Tagger*, se analizan los resultados emitidos en las evaluaciones. Descubriendo que para los sustantivos se detectan mal el 2.62% y el 8.20% para los verbos. Indicando que se pudieron alcanzar mejores resultados de haber reducido estos errores.

4.4.2.3. ANÁLISIS GENERAL DEL PROCESO DE WSD USANDO RST+FREC

En esta aproximación se fija como objetivo, intentar superar la propuesta base de RST y MFS. Para ello se ha desarrollado un método de WSD (RST+Frec) que fusiona a ambos. Esta propuesta surge al desear detectar los sentidos de las palabras más frecuentes según el contexto analizado y conseguir una mejor distinción entre sentidos. RST+Frec ha demostrado superar todas las aproximaciones no supervisadas de WSD que respectan a las competiciones tomadas como base de prueba. Se demuestra que al utilizar un tipo de información de frecuencias de sentidos con RST+Frec, se consiguen resultados siempre superiores al MFS que utilice la misma fuente. Esto infiere que de obtener fuentes más fiables de frecuencias de sentidos, al ser integrado en RST+Frec conseguirá obtener resultados superiores que el *baseline*.

4.4.3. *N-CLIQUES* COMBINADO CON *REUTERS VECTOR*

N-Cliques+RV se evalúa sobre el corpus de la tarea *English All Words* de Senseval-2 (Cotton *et al.*, 2001). Los análisis se dividen en siete experimentos, donde se busca analizar en detalle la influencia de las diferentes combinaciones de recursos en la propuesta. Luego de hacer pequeñas experimentaciones con diferentes valores de distancia *N*, se ha comprobado que la distancia entre vértices de valor 3 se ajusta a mantener un balance entre el mejor resultado en precisión y tiempo de ejecución. Entonces, se fija $N =\cdot 3$ para desarrollar las evaluaciones. Los experimentos enumerados a continuación aplican el proceso de WSD usando *N-Cliques* y para la creación del grafo inicial con *Reuters Vector*. Como rasgo común entre estas

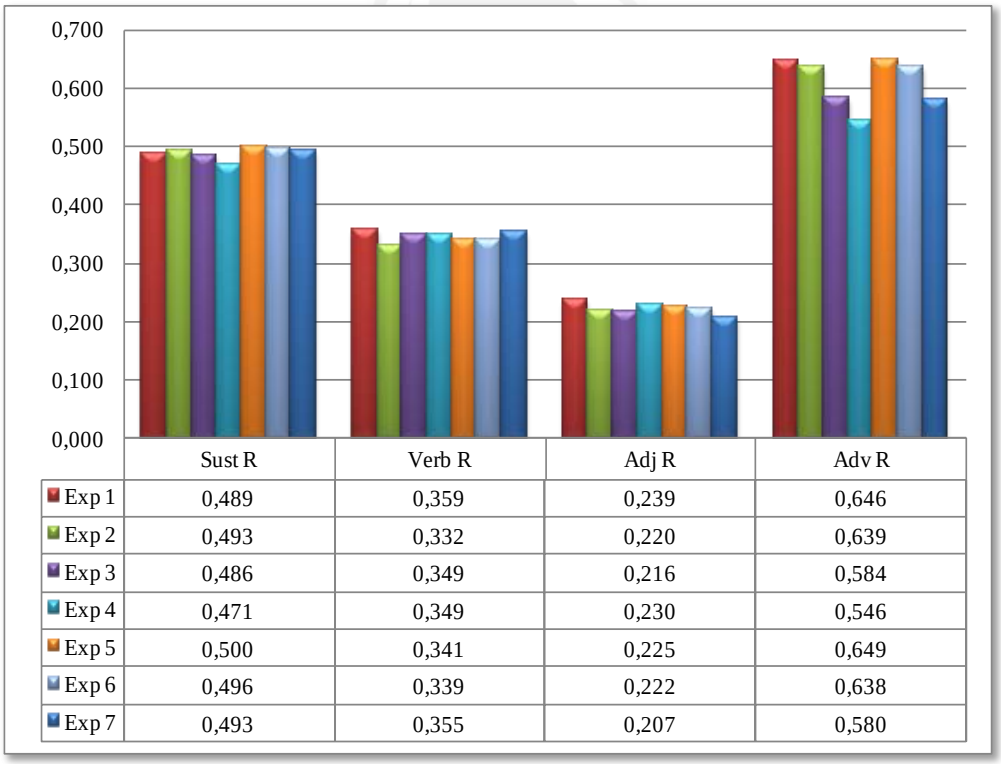
experimentaciones se puede decir que solamente varían las dimensiones que se tienen en cuenta para la creación del $G_{D.}$

- Exp 1: *N-Cliques*+RV con $G_{D.}$compuesto por WN, WND, WNA y SUMO.
- Exp 2: *N-Cliques*+RV con $G_{D.}$compuesto por WN y WND.
- Exp 3: *N-Cliques*+RV $G_{D.}$compuesto por WN y SUMO.
- Exp 4: *N-Cliques*+RV con $G_{D.}$compuesto por WN y WNA.
- Exp 5: *N-Cliques*+RV con $G_{D.}$compuesto por WN, WND y SUMO.
- Exp 6: *N-Cliques*+RV con $G_{D.}$compuesto por WN, WND y WNA.
- Exp 7: *N-Cliques*+RV con el $G_{D.}$compuesto por WN, WNA y SUMO.

Los resultados de estos experimentos se ilustran en la Figura 35 y seguidamente se exhiben varios análisis realizados sobre ellos.

4.4.3.1. ANÁLISIS DE LOS RESULTADOS PARA CADA CATEGORÍA GRAMATICAL

En cada experimento realizado se ha analizado el comportamiento de cada categoría gramatical. En la Figura 35 se observa cómo la propuesta obtiene la más alta precisión en la determinación de los adverbios, alcanzando el 64%. Los sustantivos, adjetivos y verbos obtienen baja precisión, respectivamente. Se pudiera deducir que la propuesta es muy eficaz para eliminar la ambigüedad de los adverbios si se compara con los resultados obtenidos por los sistemas en Senseval-2 (véase la sección 2.6.1.2). Relacionados con los sustantivos y adjetivos la precisión llega al 50% y 35% respectivamente, lo cual indica un comportamiento regular. Por último, los verbos obtienen los peores resultados, estos se aproximan al 25%, su determinación tan errónea se debe a que es la categoría más polisémica.



	Sust R	Verb R	Adj R	Adv R
Exp 1	0,489	0,359	0,239	0,646
Exp 2	0,493	0,332	0,220	0,639
Exp 3	0,486	0,349	0,216	0,584
Exp 4	0,471	0,349	0,230	0,546
Exp 5	0,500	0,341	0,225	0,649
Exp 6	0,496	0,339	0,222	0,638
Exp 7	0,493	0,355	0,207	0,580

Figura 35. Resultados de *N-Cliques*+RV (*Precision* (P) y *Recall* (R)).

4.4.3.2. ANÁLISIS DEL COMPORTAMIENTO DE LOS SENTIDOS CORRECTOS DE ACUERDO A LOS DIFERENTES *N-CLIQUES*

Una vez obtenidos los *N-Cliques*, cada uno contiene una serie de sentidos candidatos. Con el fin de evaluar la precisión de cada uno de *N-Clique,* se aplica una comparativa entre sentidos candidatos por el método y los que realmente son correctos acorde con Senseval-2. La Tabla 31 muestra dos tipos de información, por un lado los *N-Cliques* ordenados cantidad de sentidos propuestos como correctos incluidos en ellos y en el otro lado los *N-Cliques* con las cantidades de sentidos realmente correctos que están presentes en ellos. Los *N-Cliques* se ordenan decrecientemente con el fin de determinar cuáles son los *N-Cliques* más relevantes para un sistema de WSD.

Se puede apreciar que en la selección de los cinco *N-Cliques* que acumulan la mayor cantidad de sentidos propuestos y correctos se evidencia un 100% de coincidencia de relevancia. Es obvio que no coinciden en cantidades debido que la propuesta alcanza resultados de alrededor del 40% de *Precision* y *Recall*. Nótese, que la idea principal de la propuesta de *N-Cliques* en WSD, es lograr que los sentidos correctos de una frase objetivo se encuentren incluidos en los primeros *N-Cliques* y como se observa en la Tabla 31 el comportamiento es consecuente.

Propuesta de WSD		**Correctos según Senseval-2**	
N-Cliques	Cantidades de sentidos candidatos	*N-Cliques*	Cantidades de sentidos correctos
Clique0	361	*Clique0*	195
Clique2	268	*Clique3*	158
Clique3	251	*Clique1*	146
Clique1	251	*Clique2*	143
Clique4	229	*Clique4*	140

Tabla 31. *N-Cliques* ordenados por cantidad de sentidos candidatos a correctos y correctos.

4.4.3.3. ANÁLISIS GENERAL DE *N-CLIQUES*+RV

La integración de los recursos de WordNet en WSD, es una nueva forma de obtener resultados relevantes para enfrentar la problemática de la ambigüedad semántica. El método propuesto es una de las variantes que se puede utilizar sin lugar a dudas. Según demuestra la Tabla 32, el experimento más importante ha sido el que utiliza todos los recursos del ISR-WN en su primera versión, este obtiene un 43,3% de *Recall*. Este resultado pudiera ubicar la propuesta en el lugar onceno del *ranking* de Senseval-2 según la Tabla 32. Resultados de *N-Cliques+RV* (*Precision* (P) y *Recall* (R)).

Si se aplica el método de WSD solamente teniendo en cuenta en el G_D a WN y WNA, se obtiene una alta precisión del 69,3%, dato muy relevante para esta competición. Nótese, que en ocasiones el uso de WNA marca una diferencia notable con respecto a los demás recursos, por lo que se debe prestar cuidadosa atención a su uso. Esto indica que de poder aplicar esta propuesta sobre un corpus más afectivo (que contenga más información que denote sentimientos) la propuesta podría mejorar aumentando su cobertura. De modo general, se considera que funciona relevantemente en la detección del sentido correcto de los adverbios, y no muy precisa para los verbos. El resto, se desenvuelve regularmente.

Experiementos	Dimensiones en ISR-WN				*Recall*	*Precision*
	WN	SUMO	WND	WNA		
Exp 1	**X**	**X**	**X**	**X**	**0.433**	**0.444**
Exp 2	X		X		0.422	0.432
Exp 3	X	X			0.409	0.432
Exp 4	**X**			**X**	**0.397**	**0.693**
Exp 5	X	X	X		0.429	0.438
Exp 6	X		X	X	0.425	0.434
Exp 7	X	X		X	0.411	0.434

Tabla 32. Resultados de *N-Cliques*+RV (*Precision* (P) y *Recall* (R)).

Rank	*Precision*	*Recall*	*Sistema*		*Rank*	*Precision*	*Recall*	*Sistema*
1	0.690	0.690	SMUaw		12	0.748	0.357	IRST
2	0.636	0.636	CNTS-Antwerp		13	0.345	0.338	USM 1
3	0.618	0.618	S-LIA-HMM		14	0.336	0.336	USM 3
4	0.575	0.569	UNED-AW-U2		15	0.572	0.291	BCU
5	0.556	0.550	UNED-AW-U		16	0.440	0.200	Sheffield
6	0.475	0.454	UCLA - gchao2		17	0.566	0.169	S - sel-ospd
7	0.474	0.453	UCLA - gchao3		18	0.545	0.169	S - sel-ospd-ana
8	0.416	0.451	CLR-DIMAP		29	0.598	0.140	S - sel
9	0.451	0.451	CLR-DIMAP (R)		20	0.328	0.038	IIT 2
10	0.500	0.449	UCLA - gchao		21	0.294	0.034	IIT 3
-	**0.444**	**0.433**	**Exp 1 (Mejor)** ***N-Cliques*****+RV**		22	0.287	0.033	IIT 1
11	0.360	0.360	USM 2					

Tabla 33. Posicionamiento de *N-Cliques*+RV sobre el corpus de Senseval-2.

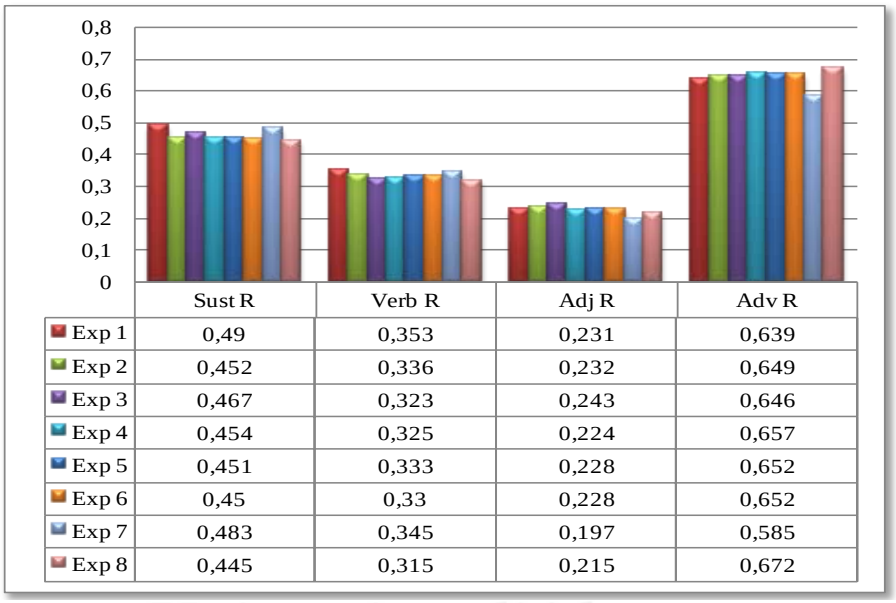
4.4.4. *N-CLIQUES* COMBINADO CON ÁRBOLES SEMÁNTICOS RELEVANTES

N-Cliques+RST se evalúa sobre corpus de Senseval-2 en particular sobre la tarea *English All Words*. Para ello el análisis se ha dividido en ocho experimentos descritos en detalle, con el objetivo de evaluar la influencia de diferentes combinaciones de recursos y determinar si la sustitución de *Reuters Vector* en la creación del grafo de desambiguación, varía radicalmente los resultados. La distancia experimental utilizada en la técnica de partición es $N = 3$. Esta distancia es más eficaz y más rápida que otras distancias superiores. Sin embargo, con distancias menores produce peores resultados. Los experimentos enumerados a continuación aplican el proceso de WSD usando *N-Cliques* y para la creación del grafo inicial se utiliza RST. Aquí se varían las dimensiones para la creación del grafo de desambiguación y las implicadas en la construcción de los árboles semánticos.

- Exp 1: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de WND.
- Exp 2: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de SUMO.
- Exp 3: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de WNA.
- Exp 4: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de WN.
- Exp 5: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de WND y SUMO.
- Exp 6: *N-Cliques*+RST con $G_{D.}$compuesto por WN, WND, WNA y SUMO, con el uso de RST de WND, SUMO, WNA y WN.
- Exp 7: *N-Cliques*+RST con $G_{D.}$compuesto por WN y SUMO, con el uso de *RST* de SUMO.
- Exp 8: *N-Cliques*+RST con $G_{D.}$compuesto por WN y WND, con el uso de *RST* de WND.

4.4.4.1. ANÁLISIS DE LOS RESULTADOS PARA CADA CATEGORÍA GRAMATICAL

El análisis de cada experimento se realiza con el fin de conocer el comportamiento de cada categoría gramatical al aplicar el proceso de WSD. Como se observa en la Figura 36, la desambiguación de adverbios obtiene de igual forma que en la propuesta con *Reuters Vector* la más alta *Precision*, llegando a 67%. Por otro lado, los verbos y los adjetivos tienen una baja *Precision*, respectivamente. Luego de analizar los resultados, esta propuesta puede ser considerada de gran alcance para eliminar la ambigüedad de los adverbios. En la distinción de sustantivos y adjetivos se alcanzan resultados de alrededor del 49% y 35% respectivamente. Por último, se evidencia que los verbos obtienen los peores resultados en torno a un 24%, debido a que es la categoría más polisémica.



	Sust R	Verb R	Adj R	Adv R
Exp 1	0,49	0,353	0,231	0,639
Exp 2	0,452	0,336	0,232	0,649
Exp 3	0,467	0,323	0,243	0,646
Exp 4	0,454	0,325	0,224	0,657
Exp 5	0,451	0,333	0,228	0,652
Exp 6	0,45	0,33	0,228	0,652
Exp 7	0,483	0,345	0,197	0,585
Exp 8	0,445	0,315	0,215	0,672

Figura 36. Resultados de *N-Cliques*+RST (*Precision* (P) y *Recall* (R)).

4.4.4.2. ANÁLISIS GENERAL DE *N-CLIQUES*+RST

En esta sección se realiza un análisis general para detectar si la integración de los recursos de WordNet ayuda a la tarea de WSD. También se desea conocer si realmente existen diferencias notables entre el uso de *Reuters Vector* y RST como propuestas para la creación del G_D.

La Tabla 34 muestra que la mayoría de los experimentos que integran todas las dimensiones, son capaces de mejorar los resultados de los experimentos 7 y 8 (estos son los que menos dimensiones incluyen), con la excepción del experimento 6. Esto indica que la técnica de particionamiento aplicada a G_D adquiere un rendimiento eficaz a medida que incrementa su información semántica. La precisión y cobertura se calcula utilizando las mismas medidas que en la competencia Senseval (descritas en la sección 2.6), en correspondencia con *Precision* y *Recall* respectivamente. El experimento más relevante es el que utiliza todos los recursos de la primera versión de ISR-WN (pero creado el grafo inicial con RST de WND), donde el *Recall* obtenido alcanza el 42,6%. Eso indica que WND es el recurso semántico más influyente entre todos (característica identificada en también en la sección de evaluación de RST como método de WSD). Este resultado podría ubicar la propuesta en el lugar décimo primero del *ranking* de Senseval-2. Es importante destacar, que el uso de WNA en la creación del RST mejora todos los resultados de precisión obtenidos por esta propuesta, característica muy similar a lo ocurrido con el uso de *Reuters Vector*.

Experiementos	Dimensiones en ISR-WN				Dimensiones en RST				*Recall*	*Precision*
	WN	SUMO	WND	WNA	WN	SUMO	WND	WNA		
Exp 1	**X**	**X**	**X**	**X**			**X**		**0.426**	**0.436**
Exp 2	X	X	X	X		X			0.407	0.416
Exp 3	**X**	**X**	**X**	**X**				**X**	**0.413**	**0.515**
Exp 4	X	X	X	X	X				0.405	0.414
Exp 5	X	X	X	X		X	X		0.405	0.414
Exp 6	X	X	X	X	X	X	X	X	0.397	0.405
Exp 7	X	X				X			0.402	0.425
Exp 8	X		X				X		0.398	0.406

Tabla 34. Resultados de *N-Cliques* usando RST (*Precision* (P) y *Recall* (R)).

4.4.5. *PERSONALIZING PAGERANK* COMBINADO CON FRECUENCIAS DE SENTIDOS BASADO MÚLTIPLES DIMENSIONES

En esta sección se desea conocer si las nuevas propuestas planteadas en la sección 4.3.2.1, que sugieren utilizar las frecuencias de sentidos en el vector de importancia *v* de la propuesta *Personalizing PageRank* sobre los nodos representativos de las frases y múltiples bases de conocimiento, logran obtener resultados relevantes en la tarea de WSD. Para conocer si es aceptada esta idea, se toman como corpus de evaluación los de la tarea *"English All Words"* de Senseval-2 de y Senseval-3. El objetivo principal de estas elecciones de corpus, es establecer comparaciones que contrasten con las novedosas propuestas basadas en grafos, que en esta Tesis se han tomado como referencia. En concreto la comparación gira alrededor de la propuestas de (Sinha and Mihalcea, 2007, Agirre and Soroa, 2009) y otras que se consideran relevantes, por ser las de mayores resultados alcanzados por sistemas de este tipo hasta la fecha.

La evaluación se divide en dos etapas según su aplicación sobre cada corpus en particular. En general en cada experimentación del Ppr+Frec, se enfatiza en conocer cómo sería su comportamiento según la LKB utilizada. Además de conocer según la LKB, si es capaz de obtener variantes especializas en clasificar bien ciertos conjuntos de palabras disjuntos entre sí. En las siguientes secciones se aborda cada una de las variantes mencionadas.

4.4.5.1. EVALUACIÓN CON EL CORPUS DE SENSEVAL-2

Los experimentos evaluados en esta sección, difieren en la base de conocimiento utilizada. Entre ellos se propone una aproximación de votación básica, que propone como sentido correcto, el que para una palabra la mayoría de las aproximaciones lo elijan como correcto. La experimentación de votación comprende a los candidatos de los experimentos 1, 2, 3, 4 y 5. En caso de no existir un sentido prevaleciente se toma como correcto el que propone el experimento 1 (MFS).

Como se puede observar en la Tabla 35, en las primeras cinco evaluaciones se toman como LKB's los recursos (o dimensiones) por individual. La primera únicamente aplica MFS y en caso de fallar toma el primer sentido para una palabra que provee WN. La segunda y tercera bases de conocimientos, ofrecen a WN nuevas relaciones de carácter léxico. ISR-WN (incluyendo WN, WND, WNA, SUMO y SC) proporciona revelaciones de enriquecimiento conceptual a WN. Y pSemcor ofrece informaciones de coocurrencia de colocación entre sentidos según el corpus de SemCor.

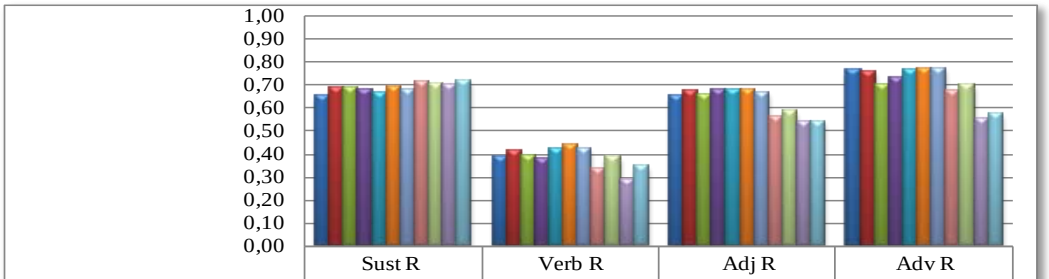
Como se puede observar, se encuentra el *baseline* MFS el cual obtiene altos resultados. Este ha sido utilizado en esta experimentación y sirve como información adicional para el uso en Ppr+Frec. Tal y como se aprecia en la Tabla 35, todas las evaluaciones de Ppr+Frec consiguen superar al *baseline* MFS. Esto indica que el haber usado las frecuencias de sentidos en el vector *v* ofrece informaciones relevantes en el proceso de WSD y hace que la nueva propuesta sea

mejor que ambas por separado. Por ejemplo, las evaluaciones que se toman de referencia (Experimentos 8, 9, 10 y 11) de (Agirre and Soroa, 2009), han sido superadas en amplio margen. Estas evaluaciones obtenidas de (Agirre and Soroa, 2009) corresponden a *Personalizad PageRank* (Ppr) y *Personalizad PageRank_w2w* (Ppr_w2w) aplicadas con LKB1.7 y LKB3.0. Es importante resaltar que según el uso de cada fuente de conocimiento en particular, existe un comportamiento variable hacia la clasificación de determinadas palabras. Esto provoca la evaluación de una votación que alcanza los resultados más relevantes entre todas las ejecuciones. El experimento número 7 obtiene buenos resultados, pero sin embargo no los imaginados, pues este ha sido aplicado teniendo en cuenta todos los recursos integrados en la LKB. Al parecer en la creación del G_D en esa evaluación, no se logra considerar toda la información necesaria para aplicar de forma efectiva el proceso de WSD. Es importante destacar que el solo hecho de asignar ponderaciones de frecuencia de sentidos en el experimento 2, se consigue superar en un 6.2% al experimento 8 (Ppr), que básicamente es igual.

Experimentos	**Recursos**					**Votación**	***Recall***	***Precision***
	MFS+ *First Sense*	LKB17	LKB30	ISR-WN	pSemcor			
Exp 1MFS	X						0.604	0.609
Exp 2		**X**					**0.630**	**0.637**
Exp 3			X				0.614	0.619
Exp 4				X			0.615	0.621
Exp 5					**X**		**0.623**	**0.628**
Exp 6						**X**	**0.641**	**0.646**
Exp 7	**X**	**X**	**X**	**X**	**X**		**0.629**	**0.634**
8- Ppr		X					0.568	
9-Ppr_w2w		X					0.586	
10- Ppr			X				0.535	
11-Ppr_w2w			X				0.558	

Tabla 35. Resultados generales de Ppr+Frec evaluado sobre Senseval-2.

Por otra parte, es posible analizar cómo ha sido el comportamiento de las diferentes categorías gramaticales a través de las evaluaciones. La Figura 37 ofrece una gráfica donde salta a la vista, que el uso del *PageRank* propuesto por Agirre para los sustantivos, supera por poco margen a las demás experimentaciones. Sin embargo, para todas las demás categorías las nuevas aproximaciones son ampliamente mejores.



	Sust R	Verb R	Adj R	Adv R
MFS 1	0,652	0,386	0,649	0,764
Exp 2	0,688	0,413	0,669	0,761
Exp 3	0,686	0,392	0,658	0,701
Exp 4	0,679	0,381	0,675	0,731
Exp 5	0,665	0,417	0,677	0,767
Exp 6	0,690	0,440	0,682	0,774
Exp 7	0,679	0,422	0,667	0,774
8-Ppr	0,711	0,334	0,559	0,671
9-Ppr_w2w	0,704	0,389	0,583	0,701
10-Ppr	0,700	0,286	0,539	0,551
11-Ppr_w2w	0,719	0,344	0,538	0,575

Figura 37. Comportamiento de categorías gramaticales con Ppr+Frec sobre Senseval-2.

La mejor aproximación de Ppr+Frec al compararla con los resultados obtenidos por los sistemas de Senseval-2 según la Tabla 36, supera a varios sistemas supervisados, pudiendo colocarse en el segundo lugar del *ranking* sin contar con el *baseline* que le antecede (este *baseline* no constituye un sistema de la competición). Esto sugiere que si el MFS computado en la aproximación propuesta, obtiene un 60,4% de *Recall* y el obtenido en el artículo de (Preiss, 2006) es mejor con un 64,6%, entonces si ya se ha demostrado que según el recurso de frecuencias de sentidos que se utilice en Ppr+Frec, se es capaz de superar en amplio margen, se genera una nueva incertidumbre. ¿Si Ppr+Frec supera en un 3.7% al *baseline* utilizado, qué exactitud podría obtener si utilizara como suministro de frecuencias la fuente del MFS de Preiss? Esta pregunta se intentará resolver cuando sea posible obtener dichas fuentes en un futuro.

Precision	***Recall***		***System***	***Supervised***
0.690	0.690		*SMUaw*	S
0.669	0.646		*Baseline-MFS-Preiss*	-
0.646	**0.641**		**Exp 6 (Ppr+Frec)**	**U**
0.636	0.636		*CNTS-Antwerp*	S
0.618	0.618		*Sinequa-LIA - HMM*	S
0.617	0.617		*Baseline-MFS-Chen*	
0.575	0.569		*UNED - AW-U2*	U
0.556	0.550		*UNED - AW-U*	U
0.475	0.454		*UCLA - gchao2*	S
0.474	0.453		*UCLA - gchao3*	S
0.416	0.451		*CL Research - DIMAP*	U
0.451	0.451		*CL Research - DIMAP (R)*	U
0.500	0.449		*UCLA - gchao*	S
0.360	0.360		*Universiti Sains Malaysia 2*	U
0.748	0.357		*IRST*	U
0.345	0.338		*Universiti Sains Malaysia 1*	U
0.336	0.336		*Universiti Sains Malaysia 3*	U
0.572	0.291		*BCU - ehu-dlist-all*	S
0.440	0.200		*Sheffield*	U
0.566	0.169		*Sussex - sel-ospd*	U
0.545	0.169		*Sussex - sel-ospd-ana*	U
0.598	0.140		*Sussex - sel*	U
0.328	0.038		*IIT 2*	U
0.294	0.034		*IIT 3*	U
0.287	0.033		*IIT 1*	U

Tabla 36. Comparación de Ppr+Frec con el *ranking* de Senseval-2 (Supervisado (S), sin supervisión (U)).

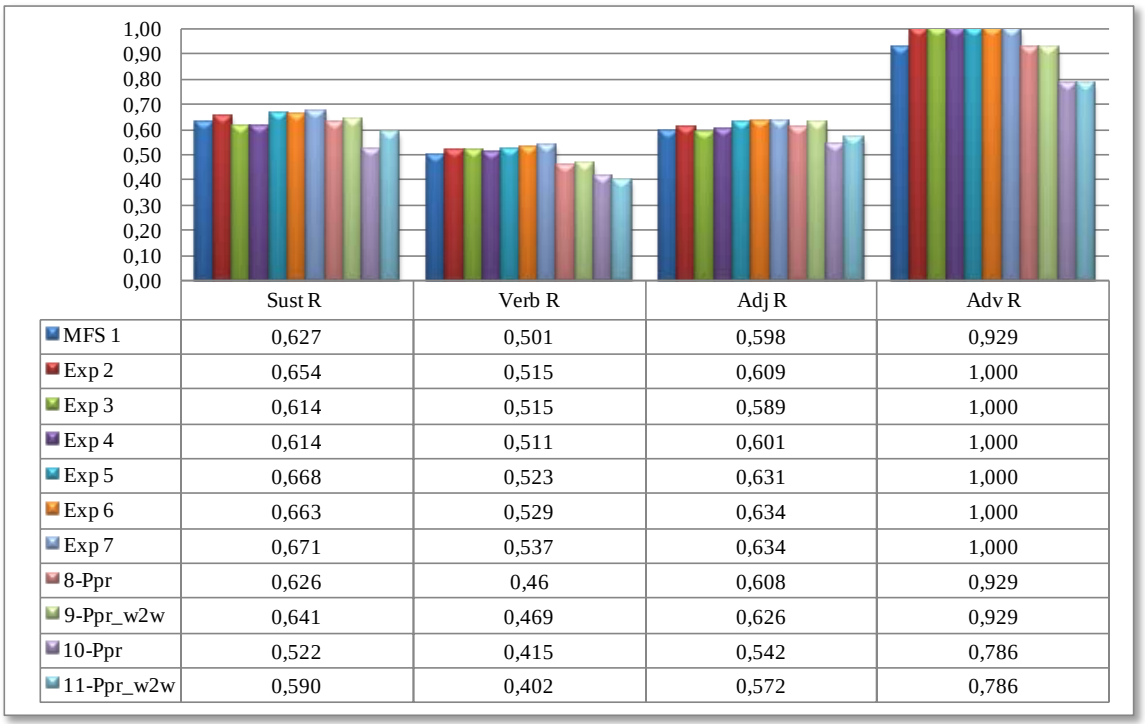
4.4.5.2. EVALUACIÓN CON EL CORPUS DE SENSEVAL-3

Las evaluaciones realizadas sobre este corpus siguen los mismos principios que sobre el corpus anterior. En la Tabla 37 se presentan varias ejecuciones capaces de superar el *baseline* MFS calculado en esta evaluación. No todas se comportan mejor, lo que demuestra la variabilidad del corpus con respecto al anterior. Es importante resaltar que las evaluaciones 5, 6 y 7 adquieren sobre ambos corpus los mejores resultados, lo que significa que los recursos y las directrices implicadas en ellos representan sin lugar a dudas elementos distintivos para WSD. Nuevamente se consigue superar los métodos de Ppr y Ppr_w2w además del *baseline*.

Experimentos	Recursos					Votación	R	P
	MFS-17 *+First Sense*	LKB17	LKB30	ISR-WN	pSemcor			
Exp 1MFS	X						0.578	0.579
Exp 2		**X**					**0.598**	**0.600**
Exp 3			X				0.576	0.578
Exp 4				X			0.576	0.578
Exp 5					**X**		**0.611**	**0.612**
Exp 6						**X**	**0.611**	**0.613**
Exp 7	**X**	**X**	**X**	**X**	**X**		**0.618**	**0.620**
8- Ppr		X					0.561	-
9-Ppr_w2w		X					0.574	-
10- Ppr			X				0.485	-
11-Ppr_w2w			X				0.516	-

Tabla 37. Resultados generales de Ppr+Frec sobre Senseval-3 (*Precision* (P) y *Recall* (R)).

La Figura 38 describe que para los adverbios no existe ni un solo fallo en Ppr+Frec, este resultado solamente es comparable con el obtenido por (Sinha and Mihalcea, 2007) la que es analizada próximamente. La mayoría de las evaluaciones de Ppr+Frec según describe la figura, superan en todas las categorías gramaticales a los métodos de referencia.



	Sust R	Verb R	Adj R	Adv R
MFS 1	0,627	0,501	0,598	0,929
Exp 2	0,654	0,515	0,609	1,000
Exp 3	0,614	0,515	0,589	1,000
Exp 4	0,614	0,511	0,601	1,000
Exp 5	0,668	0,523	0,631	1,000
Exp 6	0,663	0,529	0,634	1,000
Exp 7	0,671	0,537	0,634	1,000
8-Ppr	0,626	0,46	0,608	0,929
9-Ppr_w2w	0,641	0,469	0,626	0,929
10-Ppr	0,522	0,415	0,542	0,786
11-Ppr_w2w	0,590	0,402	0,572	0,786

Figura 38. Comportamiento de categorías gramaticales con Ppr+Frec sobre Senseval-3.

Estos resultados (experimentos 5, 6 y 7) al compararlos con los de los sistemas evaluados en Senseval-3 (véase la Tabla 38) pudieran ubicarse en el sexto puesto sin incluir el *baseline* MFS (Snyder and Palmer, 2004). Demostrando así la superación de varios sistemas supervisados y obteniendo el mejor resultado entre los que no presentan supervisión.

Rank	*System*	*Precision*	*Recall*	*Supervised*
1	*GAMBL-AW*	0.651	0.651	S
2	*Sense-Learner*	0.651	0.642	S
3	*Koc University*	0.648	0.639	S
4	*R2D2 English-All-Word*	0.626	0.626	-
-	*MFS Baseline* (*GAMBL-AW*)	0.624	0.624	
5	*Meaning All-words*	0.625	0.623	S
-	**Exp 7(Ppr+Frec)**	**0.620**	**0.618**	**U**
-	**Exp 6(Ppr+Frec)**	**0.613**	**0.611**	**U**
-	**Exp 5(Ppr+Frec)**	**0.612**	**0.611**	**U**
6	*Meaning simple*	0.611	0.610	S
-	*MFS Baseline* (*Yuret*)	0.609	0.609	
7	*LCCaw*	0.614	0.606	-
8	*upv-shmm-eaw*	0.616	0.605	-
9	*UJAEN*	0.601	0.588	S
10	*IRST-DDD-00*	0.583	0.582	U
11	*Sussex-Prob5*	0.585	0.568	-
12	*Sussex-Prob4*	0.575	0.55	-
13	*Sussex-Prob3*	0.573	0.547	-
14	*DFA-Unsup-AW*	0.557	0.546	U
15	*KUNLP-Eng-All*	0.51	0.496	U
16	*IRST-DDD-LSI*	0.661	0.496	U
17	*upv-unige-CIAOSENSO-eaw*	0.581	0.48	U
18	*merl.system3*	0.467	0.456	-
19	*upv-unige-CIAOSENSO3-eaw*	0.608	0.451	U
20	*merl-system1*	0.459	0.447	-
21	*IRST-DDD-09*	0.729	0.441	U
22	*autoPS*	0.49	0.433	U
23	*clr-04-aw*	0.506	0.431	-
24	*autoPSNVs*	0.563	0.354	U
25	*merl.system2*	0.48	0.352	-
26	*DLSI-UA-All-Nosu*	0.343	0.275	-

Tabla 38. Comparación de Ppr+Frec con el *ranking* de Senseval-3 (Supervisado (S), sin supervisión (U)).

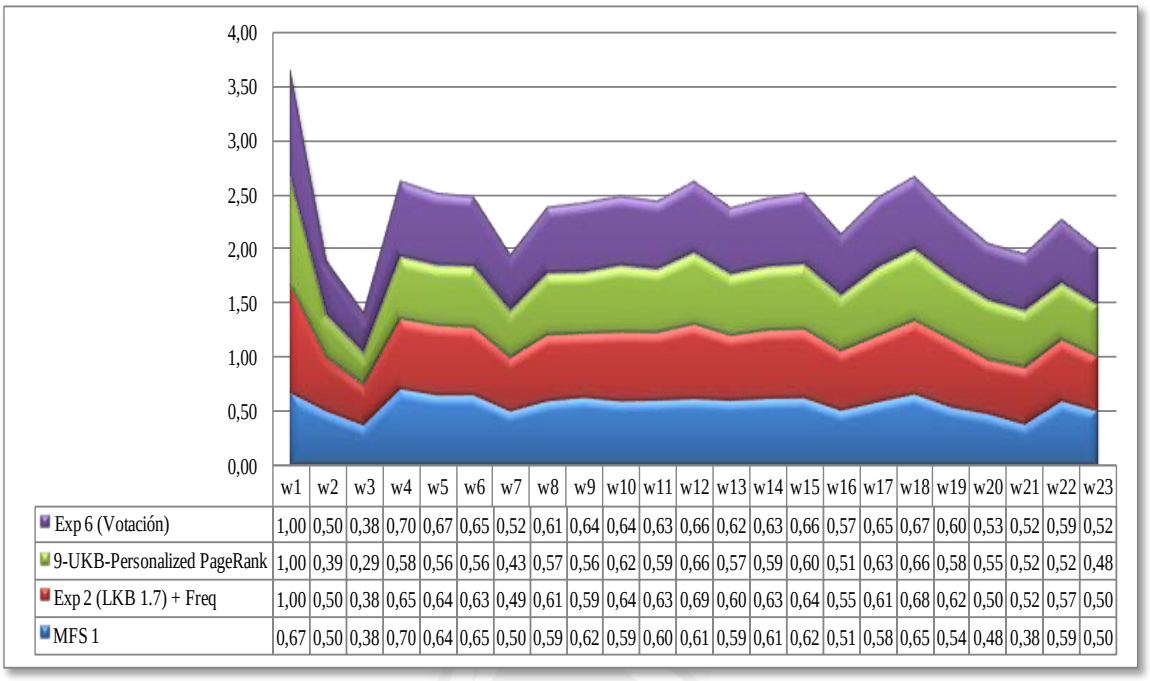
4.4.5.3. ANÁLISIS GENERAL DE PPR+FREC BASADO MÚLTIPLES DIMENSIONES

Luego de haber analizado el comportamiento de Ppr+Frec basado en múltiples dimensiones, se puede decir que está entre las aproximaciones no supervisadas basadas y en conocimiento más relevantes en comparación con las reportadas. Esta supera a todas que son de su categoría y reduce considerablemente el margen que por mucho tiempo ha existido entre sistemas supervisados y los que no lo son. Entre los aspectos más relevantes están las exactitudes obtenidas para los adverbios, en particular en la evaluación sobre el corpus de Senseval-3. Donde se logra clasificar correctamente el 100% de estos. Además, el comportamiento de las demás categorías gramaticales se mantiene cercano a los límites de los mejores resultados reportados hasta la fecha.

Al aplicar un análisis de la conducta de Ppr+Frec en las diferentes ventanas de palabras (véase la Figura 39). Se obtienen diferentes *Recalls* adquiridos para cada ventana. El grafico de área de la Figura 39 muestra como MFS, el método de votación Ppr+Frec, el Ppr+Frec usando solamente LKB1.7 y el mejor de los sistemas de (*Agirre and Soroa, 2009*) *Personalizing PageRank_w2w* únicamente usando LKB1.7, sostienen un comportamiento similar. Esto quiere decir, que las exactitudes obtenidas se mantienen reguladas por las dificultades que presenta en corpus de evaluación, como entre MFS y *Personalizing PageRank_w2w* no existe ninguna ligadura, incluso así su comportamiento frente al corpus es similar.

Sí se puede afirmar que la propuesta de Ppr+Frec está ligada al comportamiento de MFS, consiguiendo en todos los casos superar el *baseline*. En caso de obtener fuentes de información

de frecuencias de sentidos más eficaces, mejores serán los resultados de Ppr+Frec. Es importante revelar, que se han integrado e independizado a la vez múltiples dimensiones. Pero se ha de acentuar que no todas son semánticas, aquí se combina la frecuencia, la colocación (usando pSemcor) y la semántica de sentidos de las palabras (usando XWN1.7, XWN3.0, WN, WND, WNA, SUMO y SC). Con ello, se consigue establecer una interrelación capaz de abarcar la distinción de palabras de mucha o poca relación contextual.



	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	w15	w16	w17	w18	w19	w20	w21	w22	w23
Exp 6 (Votación)	1,00	0,50	0,38	0,70	0,67	0,65	0,52	0,61	0,64	0,64	0,63	0,66	0,62	0,63	0,66	0,57	0,65	0,67	0,60	0,53	0,52	0,59	0,52
9-UKB-Personalized PageRank	1,00	0,39	0,29	0,58	0,56	0,56	0,43	0,57	0,56	0,62	0,59	0,66	0,57	0,59	0,60	0,51	0,63	0,66	0,58	0,55	0,52	0,52	0,48
Exp 2 (LKB 1.7) + Freq	1,00	0,50	0,38	0,65	0,64	0,63	0,49	0,61	0,59	0,64	0,63	0,69	0,60	0,63	0,64	0,55	0,61	0,68	0,62	0,50	0,52	0,57	0,50
MFS 1	0,67	0,50	0,38	0,70	0,64	0,65	0,50	0,59	0,62	0,59	0,60	0,61	0,59	0,61	0,62	0,51	0,58	0,65	0,54	0,48	0,38	0,59	0,50

Figura 39. *Recalls* para cada ventana de palabras sobre Senseval-2.

4.4.6. COMPARACIONES CON PROPUESTAS NOVEDOSAS

En esta sección se establece un análisis comparativo entre algunos métodos de WSD que se consideran relevantes para la comunidad científica en la actualidad y las propuestas que en esta Tesis se defienden. En la Tabla 41 se describen varias aproximaciones que han sido relevantes para el estudio de la resolución de la ambigüedad semántica.

Tras analizar las aproximaciones basadas en grafos, se evidencia que la columna vertebral está constituida por la aplicación de medidas de detección de centralidad estructural basada en la asignación de pesos a las relaciones entre nodos. Como se puede ver en la Tabla 39 y Tabla 40 los resultados más bajos para este tipo de sistemas lo obtienen los *N-Cliques*. Aunque este no se encuentra entre los más relevantes, ha demostrado su decorosa aplicabilidad en WSD. Únicamente sería necesario introducir algún tipo de modificación, con tal de que su ejecución se base en la asignación de pesos entre relaciones de nodos y así lograr una mejor distinción entre sentidos. Analizando por orden según eficacia, como más relevante se tiene Ppr+Frec (se toma el de mejor resultados), este para el proceso de WSD no solamente determina centralidad, sino que tiene en consideración múltiples dimensiones en su LKB (colocación de sentidos, recursos semánticos y frecuencia de sentidos). Esta propuesta se concibe luego del análisis de todos los métodos considerados relevantes, identificando fortalezas y debilidades. Entre las fortalezas se detectan las medidas de centralidad en una red bajo la asignación de pesos de importancia estructural. Como debilidades, el poder asignar pesos iniciales de importancia entre nodos capaces de ofrecer realmente información útil al proceso de centralidad y la necesidad de incorporar múltiples dimensiones como bases de conocimiento, capaces de juzgar a la frase desde variados puntos de observación. Esta última debilidad también se tiene en cuenta para la obtención del RST+Frec. Si se observa que Mc.Carthy04 y RST+Frec se enfocan en obtener los

sentidos más relevantes según un contexto determinado, ambos obtienen resultados muy cercanos y a la vez relevantes. Estas dos propuestas se colocan por caminos que se han de tener en cuenta como pilares del WSD, pues han demostrado que el uso de frecuencias según el contexto funciona eficazmente para esta tarea.

Se puede decir entonces, que las dos propuestas líderes de esta Tesis, Ppr+Frec y RST+Frec se inician carentes de la distinción de importancia entre sentidos (ej. frecuencias de sentidos) y demostraron que superando esas debilidades los sistemas sin supervisión basados en conocimiento multidimensional son capaces de obtener los mejores resultados generales en esa categoría.

System	*Recall*	Sust R	Verb R	Adj R	Adv R
Mih05	0,542	0,575	0,365	56,7	0,709
Sihna07	0,564	0,656	0,323	0,614	0,602
Tsatsa07	0,492	-	-	-	-
Ppr w2w	0,586	0,704	0,389	0,583	0,701
Mc.Carthy04	-	0.63	-	-	-
(Mejor) RST	0,424	-	-	-	-
(Mejor) RST+Frec	**0,609**	**0.61**	-	-	-
(Mejor) N-Cliques+*RV*	0,433	0,489	0,359	0,239	0,646
(Mejor) N-Cliques+*RST*	0,426	0,49	0,353	0,231	0,639
(Mejor) Ppr+Frec	**0,641**	**0,69**	**0,44**	**0,682**	**0,774**

Tabla 39. Comparaciones de aproximaciones relevantes evaluadas sobre el corpus de *English All Words* de Senseval-2. (*Recall* (R)).

System	*Recall*	Sust R	Verb R	Adj R	Adv R
Mih05	0,522	-	-	-	-
Sihna07	0,524	0,605	0,406	0,541	100
Nav07	-	0,619	0,361	0,628	-
Ppr_w2w	0,574	0,641	0,469	0,626	0,929
Nav05	0,604	-	-	-	-
(Mejor) Ppr+Frec	**0,611**	**0,663**	**0,529**	**0,634**	**1**

Tabla 40. Comparaciones de aproximaciones relevantes evaluadas sobre el corpus de *English All Words* de Senseval-3. (*Recall* (R)).

Aproximaciones	Descripción	LKBs
Mih05 (Basado en grafos)	Presenta un algoritmo llamado *TexRank* (Mihalcea, 2005) basado en grafos construidos respecto a la secuencia de los datos de etiquetado, aplicando recorridos al azar y ponderando elementos según las dependencias entre nodos. Como resultado se crea un grafo completo ponderado formado por los synsets de las palabras en el contexto de entrada. El peso asignado en cada caso, se calcula mediante la ejecución de algoritmo de Lesk (Lesk, 1986). Una vez que el grafo es construido se aplica el algoritmo PageRank resaltando los synset más relevantes.	WN
Sihna07(Basado en grafos)	(Sinha and Mihalcea, 2007) extienden el anterior trabajo bajo la experimentación de seis medidas de similitud semántica para asignar pesos a los vínculos entre *synsets*. Además aplican cuatro medidas diferentes de centralidad para clasificar los vértices del grafo completo. Esta propuesta, al tener múltiples resultados incluye también un sistema de votación. En la creación del G_D se construye uno para cada palabra objetivo.	WN
Tsatsa07(Basado en grafos)	(Tsatsaronis et al., 2007) Aplican un proceso de dos etapas, la primera consiste en encontrar un sub-grafo acorde con el contexto realizando una búsqueda BFS sobre la LKB. Luego se procede a la ejecución de un algoritmo de activación de nodos implicando todo el sub-grafo. Esto provoca la ponderación entre vínculos de los nodos bajo el enfoque tf–idf (*term frequency–inverse document frequency* (frecuencia de términos-frecuencia de documento inversa)).	WN / WN+XWN

Ppr w2w(Basado en grafos)	(Agirre and Soroa, 2009) Plantean de modo general la construcción de una red semántica que represente el texto a analizar. Para ello construyen un sub-grafo a partir de la LKB compuesto por WN+XWN1.7 o WN+XWN3.0 o MCR16 + XWN1.6 usando BFS entre todos los synsets que representan las palabras del contexto. En (Agirre and Soroa, 2009) intervienen tres corridas. En todas se aplica el algoritmo PageRank. Pero SPr (primera aproximación) lo aplica sobre G_D obtenido a partir de la LKB sin admitir pesos iniciales sobre ningún sentido. Ppr (segunda aproximación) es similar al anterior pero modifica inicialmente la importancia de los sentidos que representan la frase analizada. Por último Ppr_w2w realiza el mismo proceso que Ppr, en este caso para cada palabra objetivo se genera un grafo de desambiguación concentrando la probabilidad inicial de los sentidos sobre las palabras que acompañantes y no en el sentido de la palabra objetivo en sí.	WN+XWN1.7 / WN+XWN3.0 / MCR16 + XWN1.6
Nav05 (Basado en grafos)	(Navigli and Velardi, 2005, Navigli and Velardi, 2004) Desarrollan un método de WSD basado en el conocimiento sobre la base de las cadenas léxicas conocido como SSI. El proceso comienza cuando se introduce una secuencia de texto, primero se identifican las palabras monosémicas, asignándolas automáticamente a su correspondiente *synset*. Luego, en un proceso iterativo consigue asociar los *synsets* que representan la interconexión más fuerte. La interconexión se calcula mediante la búsqueda de caminos en la LKB de WN, asumiendo ciertas reglas que implican patrones semánticos. Si el algoritmo no produce ninguna salida para una instancia determinada, la selección adjudicará al MFS.	WN
Nav07 (Basado en grafos)	(Navigli and Lapata, 2007) realizan un proceso de dos etapas: primero un método explora toda la LKB con el objetivo de encontrar un sub-grafo particularmente relevante respecto al contexto. Para ello se estudian diferentes algoritmos basados en centralidad de grafos y poder decidir la relevancia de los nodos en el sub-grafo. Como resultado, a cada palabra del contexto se le asigna el concepto de rango más elevado entre sus posibles sentidos. Como dato particular, este emplea un algoritmo de búsqueda primero en profundidad con un valor de 3 niveles.	WN
Mc.Carthy04 (Basado en Frecuencias de sentidos en diferentes dominios)	(Mc.Carthy et al., 2004) obtienen los sentidos más frecuentes a partir de una variedad de recursos (Reuters Corpus y el corpus de SemCor), algunos de los cuales proporcionan información de dominio. Esta propuesta consigue uno de los resultados más significativos de proporcionado por un sistema de WSD para los sustantivos.	Corpus

Tabla 41. Descripciones de sistemas relevantes evaluados sobre Senseval-2 y Senseval-3.

4.4.7. EVALUACIÓN GENERAL

Como se puede apreciar en las tablas de *rankings* de las competiciones de Senseval, las aproximaciones (Ppr+Frec y RST+Frec) que suplen las debilidades de las anteriores (RST, *N-Cliques*) son capaces de obtener los mejores resultados para sistemas sin supervisión. Como premisa, en todo momento se mantiene la base del análisis de WSD desde perspectivas semánticamente multidimensionales. En el desarrollo de los métodos se han tenido cuenta dos pilares fundamentales (Análisis Semántico Multidimensional e introducción de información de frecuencias de sentidos) para dos tipos de procesos diferentes (basados en similitud de árboles semánticos y basados en centralidad de estructuras de grafos). Ambos grupos de métodos son muy diferentes, sin embargo son capaces de obtener resultados sumamente relevantes en comparación con otras propuestas líderes al considerar los mismos elementos distintivos (multidimensionalidad y frecuencias). Como resultado de ello, se ha reducido en alguna medida los márgenes de errores presentes en la problemática de Resolución de Ambigüedad Semántica de las Palabras. Véase como en la Tabla 42, en comparación con el mejor de los resultados en competiciones de WSD, solamente existe una diferencia de un 5%. Además se logra en todos los casos en que se consideran estos pilares (ej. Tabla 42, Tabla 43 y Tabla 44) obtener el mejor resultado para la categoría se sistemas no supervisados.

Siempre que se utiliza un recurso que provee información de frecuencias de sentidos se logra superar al *baseline* MFS que este genera y a la aproximación tomada como base para la fusión entre ambas (véase Tabla 28, Tabla 29 y Tabla 35). Otro detalle importante, es que se demuestra

que el modelo de *N-Cliques* es válido para su uso en WSD, nótese que es la primera vez que se introduce en esta tarea y ha sido capaz de obtener resultados que pudieran colocarlo en el lugar decimo primero del *ranking* de Senseval-2 (véase la Tabla 32 y Tabla 34). Este en su primera demostración en la competición de Semeval-2, obtiene malos resultados, eso ha sido producto de la presión del tiempo de competición y circunstancias fatales en lectura del corpus (véase el lugar 29 de la Tabla 43).

English All Words - Fine-grained Scoring					
Rank	*Precision*	*Recall*		*System*	*Supervised*
1	0.690	0.690		*SMUaw*	S
-	0.669	0.646		*Baseline-MFS-Preiss*	-
-	**0.646**	**0.641**		**(Mejor) Ppr+Frec**	**U**
2	0.636	0.636		*CNTS-Antwerp*	S
3	0.618	0.618		*Sinequa-LIA - HMM*	S
-	0.617	0.617		*Baseline-MFS-Chen*	
-	0.610	0.609		(*Mejor*) *RST+Frec*	U
4	0.575	0.569		*UNED - AW-U2*	U
5	0.556	0.55		*UNED - AW-U*	U
6	0.475	0.454		*UCLA - gchao2*	S
7	0.474	0.453		*UCLA - gchao3*	S
8	0.416	0.451		*CL Research - DIMAP*	U
9	0.451	0.451		*CL Research - DIMAP* (*R*)	U
10	0.5	0.449		*UCLA - gchao*	S
-	**0.444**	**0.433**		**(Mejor) N-Cliques+*Reuters Vector***	**U**
-	**0.436**	**0.426**		**(Mejor) N-Cliques+RST**	**U**
-	**0.425**	**0.424**		**(Mejor) RST**	**U**
11	0.36	0.36		*Universiti Sains Malaysia 2*	U
12	0.748	0.357		*IRST*	U
13	0.345	0.338		*Universiti Sains Malaysia 1*	U
14	0.336	0.336		*Universiti Sains Malaysia 3*	U
15	0.572	0.291		*BCU - ehu-dlist-all*	S
16	0.44	0.2		*Sheffield*	U
17	0.566	0.169		*Sussex - sel-ospd*	U
18	0.545	0.169		*Sussex - sel-ospd-ana*	U
19	0.598	0.14		*Sussex - sel*	U
20	0.328	0.038		*IIT 2*	U
21	0.294	0.034		*IIT 3*	U
22	0.287	0.033		*IIT 1*	U

Tabla 42. Lugares en el *ranking* de Senseval-2 donde se colocarían las propuestas de WSD de la Tesis.

Rank	***System***	**P**	**R**	***Supervised***
1	*GAMBL-AW*	0.651	0.651	S
2	*Sense-Learner*	0.651	0.642	S
3	*Koc University*	0.648	0.639	S
4	*R2D2 English-All-Word*	0.626	0.626	-
-	*MFS Baseline* (*GAMBL-AW*)	0.624	0.624	-
5	*Meaning All-words*	0.625	0.623	S
-	**(Mejor) Ppr+Frec**	**0.613**	**0.611**	**U**
6	*Meaning simple*	0.611	0.61	S
-	*MFS Baseline* (*Yuret*)	0.609	0.609	-
7	*LCCaw*	0.614	0.606	-
8	*upv-shmm-eaw*	0.616	0.605	-
9	*UJAEN*	0.601	0.588	S
10	*IRST-DDD-00*	0.583	0.582	U
11	*Sussex-Prob5*	0.585	0.568	-
12	*Sussex-Prob4*	0.575	0.55	-
13	*Sussex-Prob3*	0.573	0.547	-
14	*DFA-Unsup-AW*	0.557	0.546	U
15	*KUNLP-Eng-All*	0.51	0.496	U
16	*IRST-DDD-LSI*	0.661	0.496	U
17	*upv-unige-CIAOSENSO-eaw*	0.581	0.48	U

18	*merl.system3*	0.467	0.456	-
19	*upv-unige-CIAOSENSO3-eaw*	0.608	0.451	U
20	*merl-system1*	0.459	0.447	-
21	*IRST-DDD-09*	0.729	0.441	U
22	*autoPS*	0.49	0.433	U
23	*clr-04-aw*	0.506	0.431	-
24	*autoPSNVs*	0.563	0.354	U
25	*merl.system2*	0.48	0.352	-
26	*DLSI-UA-All-Nosu*	0.343	0.275	-

Tabla 43. Lugares en el *ranking* de Senseval-3 *English All Words* donde se colocarían las propuestas de WSD de la Tesis.

Rank	***System***	***Type***	***Precision***	***Recall***
1	*CFILT-2*	WS	0.570	0.555
2	*CFILT-1*	WS	0.554	0.540
3	*IIITH1-d.l.ppr.05*	WS	0.534	0.528
4	*IIITH2-d.r.l.ppr.05*	WS	0.522	0.516
-	**(Mejor) RST+Frec**	**KB**	**0,527**	**0,515**
5	*BLC20SemcorBackground*	S	0.513	0.513
-	*MFS baseline*	-	0.505	0.505
6	*BLC20Semcor*	S	0.505	0.505
7	*CFILT-3*	KB	0.512	0.495
8	*Treematch*	KB	0.506	0.493
9	*Treematch-2*	KB	0.504	0.491
10	*kyoto-2*	KB	0.481	0.481
11	*Treematch-3*	KB	0.492	0.479
12	*RACAI-MFS*	KB	0.461	0.460
13	*UCF-WS*	KB	0.447	0.441
14	*HIT-CIR-DMFS-1.ans*	KB	0.436	0.435
15	*UCF-WS-domain*	KB	0.440	0.434
16	*IIITH2-d.r.l.baseline.05*	KB	0.496	0.433
17	*IIITH1-d.l.baseline.05*	KB	0.498	0.432
18	*RACAI-2MFS*	KB	0.433	0.431
19	*IIITH1-d.l.ppv.05*	KB	0.426	0.425
20	*IIITH2-d.r.l.ppv.05*	KB	0.424	0.422
21	*UCF-WS-domain.noPropers*	KB	0.437	0.392
22	*kyoto-1*	KB	0.384	0.384
23	*BLC20Background*	S	0.380	0.380
24	*NLEL-WSD-PDB*	WS	0.381	0.356
25	*RACAI-Lexical-Chains*	KB	0.351	0.350
26	*NLEL-WSD*	WS	0.370	0.345
27	***Relevant Semantic Trees***	**KB**	**0.328**	**0.322**
28	*Relevant Semantic Trees-2*	KB	0.321	0.315
29	*Relevant Cliques*	KB	0.312	0.303
-	*Random baseline*	-	0.23	0.23

Tabla 44. Lugares en el *ranking* de Semeval-2 *English All Words on Specific Domain* donde se colocarían las propuestas de WSD de la Tesis.

4.5. CONCLUSIONES

Como resultado de la investigación realizada en este capítulo, se han desarrollado métodos informáticos no supervisados y basados en conocimiento que aplican el Análisis Semántico Multidimensional en la tarea de Resolución de la Ambigüedad Semántica de las Palabras. Se ha comenzado por introducir el análisis desde múltiples dimensiones semánticas, en concreto con RST. La primera de las propuestas de RST, es parte de un sistema participante en una competición internacional. Sus resultados demuestran la viabilidad de la idea principal, al interrelacionar múltiples recursos y resolver la ambigüedad de las palabras. Para introducir mejoras se hizo necesario el uso de otro tipo de información que fuese capaz de no solamente conceptualizar las frases, sino que también admita las relevancias existentes entre sentidos de una misma palabra. Motivados por (Mc.Carthy *et al.*, 2004) donde propone obtener los sentidos más frecuentes según el contexto donde se utilice la palabra objetivo, se genera la aproximación RST+Frec. Esta aproximación utiliza los múltiples conceptos semánticos obtenidos por RST y aplica una medida que es capaz de evaluar los valores de frecuencias de sentidos conjuntamente con los de relevancia de RST. La nueva variante, logra superar ampliamente a la original y al *baseline* MFS asociado a la frecuencia utilizada, además, obtiene los mejores resultados para un sistema de desambiguación sin supervisión, solamente superada por otra de las propuestas de esta Tesis.

Luego de lograr la interrelación de múltiples dimensiones semánticas bajo el uso de varios recursos semántico-léxicos, se genera una gran red de conocimiento ávida de ser soporte de algoritmos estructurales. Motivados por (Agirre and Soroa, 2009, Mihalcea, 2005, Navigli and Lapata, 2007, Navigli and Velardi, 2005, Sinha and Mihalcea, 2007) y otros, surge la idea de aplicar métodos de WSD basados en grafos de conocimiento. Inicialmente se propone el modelo de *N-Cliques* (Luce, 1950) combinado electivamente con RST y *Reuters Vector*. Este se desarrolla con el fin de obtener sub-grafos agrupadores de conceptos más fuertemente enlazados que representan un texto. Las evaluaciones obtenidas al aplicar esta idea, consiguen colocarse entre las posiciones medianas con respecto al *ranking* de Senseval-2. Aunque se han propuesto dos variantes (*N-Cliques*+RV y *N-Cliques*+RST) enfocadas a la creación del grafo de desambiguación, se evidencia que el problema no radica en ese punto. Solamente con establecer una red implicando todos los caminos entre los sentidos de las palabras del texto, se consigue una muestra representativa de la frase analizada. Una comparativa con otras propuestas basadas en grafos (véase Tabla 39), revela que todas las aproximaciones que aplican medidas de centralidad asociando valores de importancia a los elementos del grafo, superan a las evaluaciones de *N-Cliques*. Por esa razón, se propone una aproximación basada en grafos y multidimensional como *N-Cliques*, pero aplicando otro tipo de medida de centralidad. Se permitiría de esta forma sopesar los diferentes sentidos de las palabras asociados a sus frecuencias de aparición en los corpus. Basado en *Personalizing PageRank* (Agirre and Soroa, 2009), propuesta muy relevante pero carente esencialmente de multidimensionalidad y frecuencia de sentidos, se genera Ppr+Frec. Con esta nueva aproximación se reemplazan las insuficiencias de *Personalizing PageRank*, logrando obtener los mejores resultados entre los reportados para sistemas sin supervisión. Entre las dimensiones implicadas que mejor responden a la resolución de ambigüedad del sentido de las palabras, pSemcor y LKB1.7 siempre ofrecen respuestas muy próximas a los más relevantes, incluso se puede decir que excluyendo los reportes de la Tesis, estas ostentan los mejores de los aciertos entre los sistemas sin supervisión.

5. APLICACIONES DEL ANÁLISIS SEMÁNTICO MULTIDIMENSIONAL EN OTRAS TAREAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL

La ambigüedad del lenguaje se manifiesta de muchas maneras, esta surge cuando se necesita decidir entre varias opciones posibles. Por ejemplo, en el caso de querer clasificar un texto según su polaridad (Positivo, Negativo, Neutral), es necesario elegir entre estas tres opciones, con lo que podría considerarse necesario un proceso de desambiguación. Por estas razones, otras áreas del PLN han aplicado técnicas de WSD para resolver ciertos problemas. En este capítulo se plantea aplicar el Análisis Semántico Multidimensional a la Minería de Opiniones. A continuación se exponen diferentes aproximaciones concernientes a la Minería de Opiniones para tratar la problemática mencionada. Después se desarrollan las propuestas presentadas en esta Tesis. Finalmente, se realizan evaluaciones que conducen a las conclusiones del capítulo.

5.1. INTRODUCCIÓN

En los últimos años, la información textual se ha convertido en una de las más importantes fuentes de conocimiento para extraer datos útiles y heterogéneos. Los textos pueden proveer información sobre hechos, tales como descripciones, listas de características o instrucciones de información de opinión como revisiones, emociones o sentimientos. Esta heterogeneidad ha motivado que la comunidad científica se ocupe de la identificación, extracción de opiniones y sentimientos en los textos que requieren una atención especial y con esto, ayudar con el desarrollo de diferentes herramientas, a analistas de información del gobierno, empresas, partidos políticos, economistas, etc. Mediante este tipo de sistemas se posibilita la obtención automática de los sentimientos en las noticias y los foros, que hoy en día constituyen una gran fuente de información, siendo esta una tarea muy difícil (Wiebe *et al.*, 2005). Muchos investigadores como (Balahur et al., 2010, Hatzivassiloglou et al., 2000, Kim and Hovy, 2006, Wiebe et al., 2005) y otros, han trabajado en esa dirección. Relacionados con el área de detección de opiniones, en los últimos años los investigadores se han concentrado en el desarrollo de sistemas de respuestas a preguntas de opinión (*Opinion Question Answering* (OQA)) (Balahur *et al.*, 2010). Esta nueva tarea tiene que lidiar con diversos problemas, tales como análisis de emociones, donde los documentos deben ser clasificados de acuerdo a los sentimientos y características de la subjetividad. Para ello, se necesita un nuevo tipo de evaluación que tenga en cuenta estos novedosos aspectos.

Una de las competiciones que establece el punto de referencia para los sistemas de respuestas a preguntas de opinión, en un entorno monolingüe y multilingüe, es el NTCIR (*Multilingual Opinion Analysis Task* (MOAT[75])). En esta competición, los investigadores trabajan arduamente para lograr mejores resultados en el análisis de la opinión, con la introducción de diferentes técnicas de resolución.

Entre los distintos sistemas desarrollados existen diferentes aproximaciones. Algunos investigadores se centran en la idea de que los adjetivos combinados con características, semánticas proporcionan información esencial para el funcionamiento de análisis de la opinión (Hatzivassiloglou *et al.*, 2000). Otros, como (Zubaryeva and Savoy, 2010) asumen que la extracción de los términos relevantes en los documentos, podrían definir su polaridad, teniendo

[75] http://research.nii.ac.jp/ntcir/ntcir-ws8/meeting/

en cuenta que estos formen parte del diseño de un método capaz de seleccionar los términos que claramente pertenecen a un tipo de polaridad. Otra investigación basada en la extracción de características se ha aplicado en (Lai *et al.*, 2010), donde se desarrolla un sistema de aprendizaje sobre frases de opiniones del lenguaje japonés. Por último y no por dejar de existir casos, (Balahur and Montoyo, 2009) han propuesto un método para extraer, clasificar y resumir opiniones sobre los productos a partir de las revisiones de webs. Esta interesante propuesta, se basa en la construcción previa de taxonomías de características de productos y relaciones semánticas obtenidas por la ejecución de la distancia normalizada Google (Cilibrasi and Vitányi, 2007) y el aprendizaje SVM. Como se puede apreciar, el uso de la extracción de características es una forma muy extendida de trabajar en el área del Análisis de la Opinión. Algunos autores han utilizado recursos semánticos con este propósito, por ejemplo, (Kim and Hovy, 2006, Kim and Hovy, 2005) emplean recursos semánticos para detectar encabezados y otras tareas de la extracción de opiniones.

En la sección 2.5 donde se comentan los métodos de clasificación, se menciona que WSD constituye también un método de clasificación según la perspectiva con que se analice. Se puede decir entonces, que la tarea de establecer la polaridad de un texto, o conocer si constituye o no este texto un opinión, también puede ser vista como un tipo de clasificación de texto (Pang *et al.*, 2002). Resolver estos problemas establece de algún modo, casos especiales para tratar la ambigüedad del lenguaje humano.

Otra de las problemáticas presentes, es la capacidad de poder agrupar opiniones por categorías para posteriormente ser analizadas. Nótese que por lo general estas provienen de los foros o blogs de la Web, donde su peculiaridad es que son muy informales y su longitud textual muy corta. Esto sitúa a los sistemas de clasificación de textos en una posición poco ventajosa. Un sistema informático debe ser capaz de captar los rasgos que distinguen a cada categoría de las demás y asociar a ellas, aquellos documentos con características similares. En particular, la categorización de texto (*Text Categorization* (TC)) ha sido ampliamente estudiada por muchísimos autores (Schneider, 2005, Schütze, 1997, Diederich et al., 2003, Sebastiani, 1999, Thorsten, 1998) y ha alcanzado resultados sorprendentes al distinguir el tema (o tópico).

El enfoque básico consiste en construir un patrón representativo de cada una de las clases o categorías y los documentos. Posteriormente se aplica alguna función que permita estimar el parecido entre el documento y categorías, y asociar los más semejantes.

Existen diferentes técnicas basadas en modelos de representación vectoriales[76] que tratan este tema. Estas resultan muy sencillas y descansan sobre la premisa de que el significado de un documento puede derivarse del conjunto de rasgos presentes en él, que son analizados y evaluados según su importancia. Esto significa, que los documentos pasan a ser representados como vectores dentro de un espacio Euclidiano, de modo que al medir la distancia entre dos vectores se estima la similitud como indicador de cercanía semántica.

La primera acción a realizar es la representación adecuada de los documentos. Entre los principales modelos empleados se tiene el Modelo de Espacio Vectorial (Thorsten, 1998, Diederich et al., 2003), donde se aplican varios de los métodos de aprendizaje descritos en la sección 2.5.1. Este es un modelo matemático que ha sido usado para la representación de textos (Salton and McGill, 1986). Se caracteriza, fundamentalmente, por asumir el "principio de independencia", al considerar que las cadenas dentro en un mismo texto no tienen relación entre sí. Esta característica permite que las cadenas sean cuantificadas individualmente y además sin tener en cuenta el orden en el que aparecen en el texto. De este modo, la semántica de un documento queda reducida a la suma de los significados de los rasgos que contiene. Estas

[76] http://www.scholarpedia.org/article/Text_categorization

suposiciones, reducen drásticamente la complejidad computacional del problema, ya que permiten representar el documento simplemente como un vector.

Otro modo de clasificar textos pero sí obteniendo información semántica es utilizando LSA. El índice de latencia semántica se ha planteado como una variante al Modelo de Espacio Vectorial. Este permite establecer comparaciones de similitudes semánticas entre textos (Landauer *et al.*, 1998). Su característica fundamental (la dependencia semántica entre rasgos) es al mismo tiempo la principal diferencia con el Modelo de Espacio Vectorial. En este caso, un texto se representa en un espacio de coordenadas donde los documentos y los rasgos se expresan como una combinación de factores semánticos subyacentes.

Uno de los problemas centrales en la clasificación de texto, es la reducida dimensionalidad del espacio de características del texto. Las técnicas de clasificación estándar no pueden tratar con tales conjuntos. Cuando el contenido de un texto a categorizar es poco descriptivo, se reduce en gran medida la semántica de los documentos. Lo que sostiene entonces la afirmación de (Li and Sun, 2007) que a medida que aumenta el número de términos en el vector de características disminuye el error de clasificación. Por esta razón, retomando la tarea de análisis de opiniones en conjunto con TC, resulta sumamente relevante generar nuevas aproximaciones que sean capaces de analizar segmentos de texto (con bolsas de palabras de longitud tal donde las técnicas convencionales de TC fallan) que contengas opiniones.

En este capítulo se proponen dos nuevas aproximaciones que utilizan las propuestas de Análisis Semántico Multidimensional descritas en el capítulo anterior. El objetivo es clasificar opiniones "como la clasificación de positivos o negativos en un conjunto de textos" (Pang *et al.*, 2002), además de conocer la existencia o no de texto que contenga opiniones. Y se introduce una segunda aproximación de TC para clasificar opiniones en categorías o tópicos. El capítulo se presenta de la siguiente forma:

- Procesamiento de opiniones basado en el análisis de múltiples dimensiones semánticas.
- Clasificación de textos basado en el análisis de múltiples dimensiones semánticas.

5.2. PROCESAMIENTO DE OPINIONES BASADO EN EL ANÁLISIS DE MÚLTIPLES DIMENSIONES SEMÁNTICAS

En esta sección se propone un método sin supervisión, basado en el conocimiento suministrado por ISR-WN, a través del uso de la técnica RST combinada con SentiWordNet 3.0 (Esuli and Sebastiani, 2006). Este método se aplica para su evaluación a tres de las tareas monolingües del inglés que se proponen en la competición NTCIR 8 MOAT. En este enfoque como núcleo de ISR-WN se utiliza la versión de WN 2.0.

El objetivo de este método es obtener un RST de cada frase y luego asociar los RST's con los valores de la polaridad de SWN. El proceso incluye los siguientes recursos: WND, WNA, la taxonomía de WN, categorías de SUMO y SC. Para el caso particular de las SC que no presentan una estructura jerárquica, simplemente se obtienen las Clases Semánticas relevantes. Posteriormente, se determinan las polaridades recolectadas bajo cada etiqueta de cada RST obtenido de acuerdo con el texto analizado. La propuesta en cuestión toma como nombre Senti-RST (Gutiérrez *et al.*, 2011c) y en concreto consiste en cuatro pasos , los cuales se presentan en las secciones 5.2.1, 0, 5.2.3 y 5.2.4.

5.2.1. OBTENCIÓN DE RST'S DE CONCEPTOS

En esta sección se obtienen los Árboles Semánticos Relevantes del mismo modo que se ha fundamentado en la sección 4.2.2.2.1. Es importante resaltar que solamente se obtienen los árboles semánticos, los demás pasos de desambiguación de la propuesta de RST corresponden a la resolución de otra tarea diferente a la detección de opiniones, por tanto, se obvian.

5.2.2. OBTENCIÓN DE ÁRBOLES SEMÁNTICOS DE POLARIDAD POSITIVA

Con el fin de obtener los Árboles Semánticos Positivos (*Positive Semantic Trees* (PST)) del texto analizado, se sigue el mismo proceso descrito en la sección 4.2.2.2.1. En este caso, los valores de *AR* se sustituyen por los valores de polaridades relacionados con el sentido analizado. La polaridad se obtiene a partir del recurso SWN, donde se asigna a cada sentido de WN en ISR-WN para la versión 2.0 una etiqueta de características de SWN para la versión 3.0 de WN. Esto significa que se puede encontrar cada sentido que le da ISR-WN a las palabras, en SentiWordNet 3.0 y obtener las respectivas polaridades. Este nuevo valor de polaridad positiva asociada ahora a las etiquetas del RST obtenido, se llama Asociación Positiva (*Positive Association* ($PosA$)). El valor $PosA$ se calcula utilizando la ecuación (47).

$$PosA(C, f) = \sum_{i=1}^{n} PosA(C, f_i) \tag{47}$$

Siendo:

$$PosA(C, w) = \sum_{i=1}^{n} PosA(C, w_i) \tag{48}$$

Donde:

- C es un Concepto;
- f es una frase o conjunto de palabras (w);
- f_i es la i-ésima palabra de f;
- $PosA$ (C, w_i) representa el valor positivo del sentido (w_i) asociado al concepto C.

$PosA$ se utiliza para medir el valor positivo asociado a las hojas de los árboles semánticos en donde los conceptos se colocan. Posteriormente utilizando la misma estructura del RST, se crean nuevos árboles semánticos sin valores AR. En cambio, las hojas con los conceptos de este nuevo árbol se anotan con el valor de $PosA$.

Luego, para asignar un valor positivo a los conceptos padres, los padres de cada concepto acumulan los valores positivos de los conceptos hijos. La ecuación (49) muestra el proceso *Bottom-Up* (de abajo hacia arriba).

$$PosA(PC) = \sum_{i=1}^{n} PosA(ChC) \tag{49}$$

Donde:

- PC es el Concepto Padre
- ChC es el Concepto hijo de PC
- $PosA(ChC)$ representa el valor positivo de ChC.

5.2.3. OBTENCIÓN DE ÁRBOLES SEMÁNTICOS DE POLARIDAD NEGATIVA

En esta fase, se repite el paso descrito en la sección anterior, pero en este caso asumiendo los valores negativos de polaridades asociados. Como resultado se obtienen los Árboles Semánticos Negativos (*Negative Semantic Trees* (*NST*)) correspondiente a cada recurso semántico utilizado. Tomando como ejemplo la frase utilizada en la sección 4.2.2.2.1 donde se genera un RST de dominios, se muestra en la Tabla 45 el resultado con los valores de $PosA$ y $NegA$ (*Negative Association*) ya calculados.

Vector $AR - Pos - Neg$			
AR	*PosA*	*NegA*	*Domain*
1.63	0.00	1.00	*Social_Science*
0.90	0. 00	0.00	*Administration*
0.90	0.00	0.00	*Pedagogy*
0.80	0.00	0.00	*Root_Domain*
0.36	0.00	0.00	*Psychoanalysis*
0.36	0.00	0.50	*Economy*
0.36	0.375	0.375	*Quality*
0.36	0.00	0.00	*Politics*
0.36	0.00	0.00	*Buildings*
0.36	0.00	0.50	*Commerce*
0.36	0.00	0.00	*Environment*
0.11	0.375	0.375	*Factotum*
0.11	0.00	0.00	*Psychology*
0.11	0.00	0.00	*Architecture*
0.11	0.00	0.00	*Pure_Science*

Tabla 45. Representaciones de los PST y NST ordenados por relevancia de *AR* de los Dominios.

Según describe la Tabla 45, la frase analizada se encuentra más vinculada al dominio *Social_Science* y acumula un valor negativo uno y un valor positivo cero. Esto indica de manera superficial, que la frase es más negativa que positiva. Según los códigos de colores ilustrados en la Figura 40 y Figura 41 obtenidos por los valores representados en la Tabla 46 (según el valor de las multiplicaciones $PosA\ x\ AR$ y $NegA\ x\ AR$ se asocian colores con intensidad), se evidencia que la frase tomada como ejemplo, se sitúa en el contexto de las Ciencias Sociales prevaleciendo a grandes rasgos los sentimientos negativos. A continuación se describe la fase final con el objetivo de formalizar la decisión de polaridad de las frases, teniendo en cuenta los criterios emitidos por los árboles semánticos.

Domain	***PosA x AR***	***NegAx AR***
Social_Science	0	1.63
Administration	0	0
Pedagogy	0	0
Root_Domain	0	0
Psychoanalysis	0	0
Economy	0	0.18
Quality	0.135	0.135
Politics	0	0
Buildings	0	0
Commerce	0	0.18
Environment	0	0
Factotum	0.04125	0.04125
Psychology	0	0
Architecture	0	0
Pure_Science	0	0

Tabla 46. Valores de polaridad multiplicados relevancia para cada concepto de Dominio.

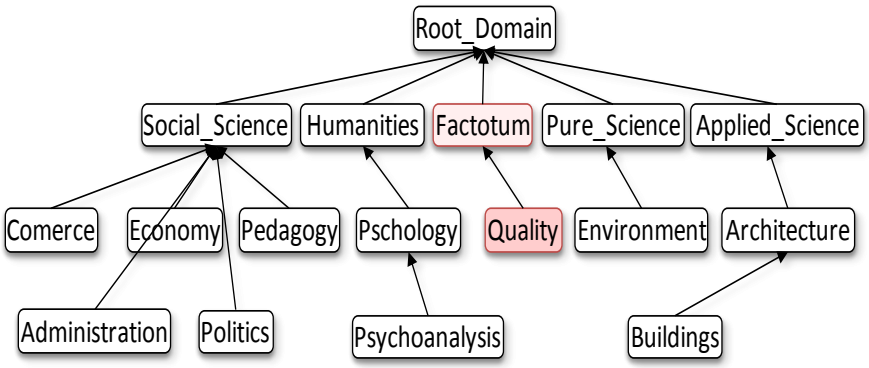


Figura 40. Árbol Relevante Positivo.

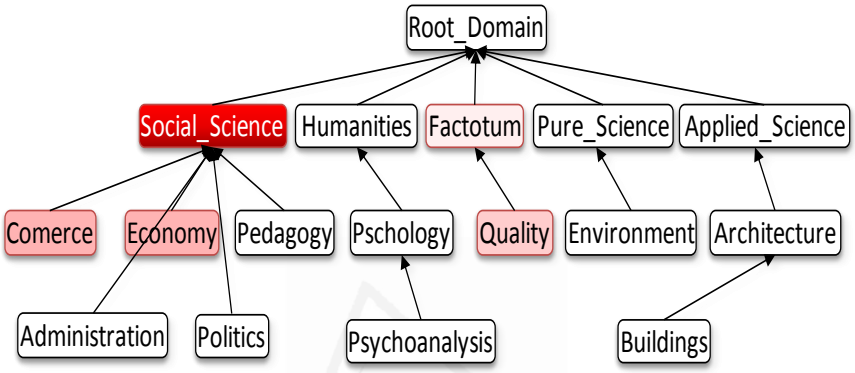


Figura 41. Árbol Relevante Negativo.

5.2.4. OBTENCIÓN DE POLARIDADES DE LAS FRASES

Este paso, se concentra en la detección de la polaridad más representativa de acuerdo con la semántica de los árboles obtenidos para cada uno de los recursos (dimensión). Para ello, se combina RST con PST y RST con NST con el fin de establecer un balance entre polaridades y relevancia de conceptos. Dependiendo de los resultados obtenidos se pueden clasificar las frases como positiva, negativa o neutral. Antes de realizar este paso, se normalizan los valores de los tres tipos de árboles (RST, PST y NST) para cada dimensión, con el objetivo de trabajar con valores entre cero y uno. Nótese que al hacer esta normalización pueden existir casos de análisis que se vean perjudicados. Al seguir el siguiente ejemplo se resaltan las dificultades que se pueden presentar en el proceso de normalización. Para los siguientes vectores obtenidos:

- *Positive* ={*Sport*|1; *Process*| 0; *Medicine*| 0}
- *Negative* {*Sport*|**0.15**; *Process* | 0; *Medicine*| 0 }

Luego de ser normalizados utilizando la ecuación (50) se obtienen las siguientes modificaciones:

- *Positive* {*Sport*|1; *Process*|0; *Medicine*|0 }
- *Negative* {*Sport*|**1**; *Process* |0; *Medicine*|0}

$$Norm_i = \frac{ST_i}{\sum_{j=1} ST_j} \tag{50}$$

Donde *ST* puede ser indistintamente un PST o NST en dependencia del que se trabaje en ese instante.

Como se puede observar, si el vector presenta pocos conceptos evaluados diferentes a cero, algunos conceptos pueden tomar valores incorrectos. Por ejemplo, en el vector negativo el valor del dominio *Sport* toma valor uno, siendo este un valor muy superior que el original de 0.15. Este hecho introduce errores en los resultados convirtiendo el vector negativo en igualdad de condiciones con respecto al vector positivo. Este resultado claramente desestima los valores de polaridades obtenidos, por lo que, este aspecto ha de tenerse en cuenta en las etapas de experimentaciones.

El principal objetivo en esta etapa es el de asignar mayor peso a las polaridades asociadas con los conceptos más relevantes en cada árbol semántico pertinente. La ecuación (51) muestra los pasos a seguir para obtener el Valor Semántico Positivo ($ACPos$) de la frase.

$$ACPosA(RST, PST) = \sum_{i=1} RST_i * PST_i \tag{51}$$

Donde:

- $ACPosA$ es el Valor Semántico Positivo Acumulado de la frase analizada respecto a una dimensión
- RST es el Árbol Semántico Relevante ordenado descendientemente con formato: RST [*Concepto*| *valor AR*]
- PST es el Árbol Semántico Positivo ordenado de acuerdo con la estructura de RST , con formato: PST [*Concepto*| *valor PosA*]
- RST_i es el i-ésimo valor de AR del concepto *i*
- PST_i es el i-ésimo valor de $PosA$ del concepto *i*

Para medir el Valor Semántico Negativo Acumulado ($ACNegA$), se emplea una ecuación similar reemplazando PST con NST. A continuación se exponen las tareas de análisis de opiniones o análisis de sentimientos por la que se rige esta propuesta. NTCIR 8 MOAT presenta varias tareas de las que se toman solamente las tres que siguen a continuación (para la configuración monolingüe del inglés):

5.2.5. DETECCIÓN DE FRASES QUE CONTIENEN OPINIONES

El enjuiciamiento de frases opinantes que contienen opiniones (*Judging sentence opinionatedness*) requiere que a los sistemas que asignen los valores de SI o NO a cada una de las oraciones en la colección de documentos en caso que cada una en particular contenga estados de opinión o no. Si la frase contiene una opinión o no, se le asignan las etiquetas "Y" o "N" correspondientemente. Para resolver esta tarea, Senti-RST analiza el PST y el NST de todas las dimensiones (WN, WSD, WNA, SUMO y SC). Después de revisar los PST's y NST's si al menos un concepto ha asignado un valor distinto de cero, el resultado es "Y" en otros casos es "N". Esta idea se basa en el principio de que, siempre que exista algún tipo de polaridad en la frase por consiguiente existe opinión.

5.2.6. DETERMINACIÓN DE FRASES RELEVANTES A PREGUNTAS

En la tarea de determinación de relevancia en las oraciones (*Determining sentence relevance*), los sistemas tienen que decidir si una frase es relevante con respecto a una pregunta y emitir un sí o un no ($Y \mid N$). Se plantea entonces que la pregunta dada se relaciona con cada frase de cada tema, si existen al menos un 50% de coincidencias entre los RST de la pregunta y la frase analizada (la similitud se obtiene por la cantidad de etiquetas de conceptos que coinciden). Además la frase analizada solamente es relevante si el PST y el NST entre todas las dimensiones contienen al menos un valor positivo o negativo.

5.2.7. CLASIFICACIÓN DE LA POLARIDAD DE LAS FRASES

En la tarea de obtención de polaridad de las frases y del tema (*Polarity and topic-polarity classification*), los sistemas deben asignar un valor de "POS", "NEG" o "NEU" (positivo, negativo o neutro) a cada una de las frases en los documentos aportados. Para ello, Senti-RST aplica el proceso de obtención de valores de $ACPos$ y $ACNeg$ de todas las dimensiones y establece una comparación entre ellos. Estos valores acumulados son nombrados $ACPosD$ y $ACNegD$ respectivamente. En el caso de $ACPosD > ACNegD$ el valor asignado es "POS", si $ACPosD < ACNegD$ el valor asignado es "NEG", de lo contrario, el valor asignado es "NEU".

5.3. CLASIFICACIÓN DE TEXTOS BASADO EN EL ANÁLISIS DE MÚLTIPLES DIMENSIONES SEMÁNTICAS

Una vez comprobadas las posibilidades que ofrecen los vectores de dominios relevantes en la representación semántica de textos cortos, y atendiendo a las deficiencias que presentan los métodos de TC, con respecto a la longitud de las colecciones de características que se pueden obtener sobre los textos a clasificar, se propone aplicar varias aproximaciones de clasificación enfocadas en la clasificación de opiniones. Esto provoca la superación de dos retos, el primero es clasificar textos cortos y el segundo lograrlo sin utilizar aprendizaje automático. Cabe destacar que por lo general las opiniones encontradas en la web y en corpus de evaluación suelen ser muy cortas. Esto genera un espacio donde los métodos tradicionales no funcionan correctamente, estos requieren de representaciones vectoriales extensas, donde a medida que aumenta el número de características disminuye su error medio. La aplicabilidad de estos métodos tradicionales sobre las opiniones comunes que se hallan hoy en día resulta ineficaz, provocando la búsqueda de nuevas variantes que particularicen este tipo de tratamiento de textos.

Por esta razón, se presenta como propuesta de solución, un método en el cual se categoricen textos cortos (aproximadamente menos de 30 palabras y una media de 12 palabras), mediante la extracción de Conceptos Semánticos Relevantes. Una vez definidos los vectores de características que van a modelar el espacio semántico del contenido de los textos, se utilizan electivamente métricas de similitud o el algoritmo de agrupamiento K-Medias (Gómez-Allende, 1993). Todas estas soluciones se realizan sin necesidad de utilizar técnicas de aprendizaje automático. Al proceso general se le llama *Multidimensional Semantic Features to Text Classification* (MSF-TC) y cuenta con los siguientes pasos que además se ilustran en la Figura 42. Como base de conocimiento se tiene a ISR-WN, en concreto los conceptos de WND, SUMO, WNA, y WN.

1. Obtención de los RST de textos y categorías
2. Normalización de Vectores
3. Aplicación de técnicas clasificación (se presentan dos variantes electivas)
    - Aplicación de K-Medias mediante la distancia Euclidiana
    - Asignación directa (con variantes electivas)
        - Obtención del máximo valor de (Resta del área de similitud con el área de desigualdad entre vectores (RSD)),
        - Obtención del mínimo valor distancia Euclidiana
        - Obtención del máximo valor del Coeficiente de Correlación de Pearson,
        - Obtención del máximo valor de la Función Jaccard,
        - Obtención del máximo valor del Coseno del ángulo

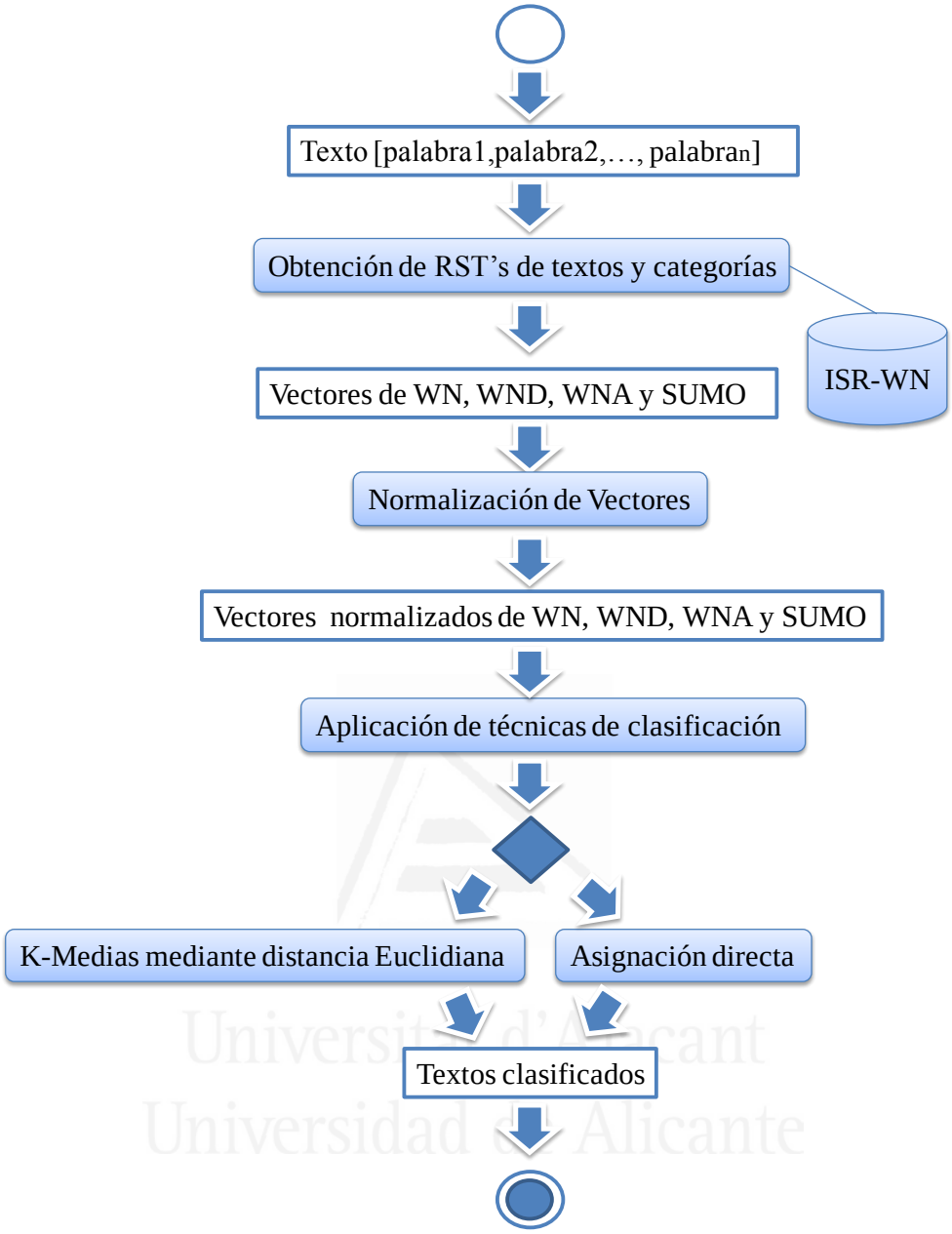


Figura 42. Método general de clasificación textual MSF-TC.

A continuación se describe cada paso del proceso, se tiene en cuenta las descripciones detalladas de las técnicas utilizadas.

5.3.1. OBTENCIÓN DE RST'S DE TEXTOS Y CATEGORÍAS

En este paso luego de tomar como entrada un texto, a este se le aplica la propuesta de obtención de Árboles Semánticos Relevantes introducida en la sección 4.2.1.2. Pero se ha de hacer la siguiente observación: con el objetivo de adaptar el proceso de obtención de Conceptos Semánticos Relevantes a la tarea de clasificación textual, el método RST no se considera en su totalidad, solamente se obtienen los conceptos relevantes sin la adición de los conceptos padres. Esta salvedad, es debida a que en la clasificación textual, en dependencia del nivel de granularidad que se desea aplicar en la clasificación, también debe tenerse en cuenta para la extracción de características. Como el campo de acción en esta ocasión son opiniones, donde varían ligeramente sus contextos, es preciso obtener solamente los conceptos relevantes

particulares y no que generalicen los textos. Como salida se obtienen un conjunto de vectores con formato [Concepto | valor AR], válidos para constituir la entrada del siguiente paso.

5.3.2. NORMALIZACIÓN DE VECTORES

Una vez obtenidos los vectores de contexto que modelan semánticamente el contenido de los textos a clasificar o textos de las categorías, el siguiente paso es utilizar una función de ponderación para normalizar los vectores. De esta manera los conceptos muy utilizados en el discurso pero que no permiten distinguir claramente los contenidos, se le reduce su valor AR en la medida que sea común o no en la colección. Es decir, al aplicarle una función de ponderación global, su valor de AR queda sujeto a la cantidad de documentos en que aparezcan dichos conceptos.

En MSF-TC se utiliza la función **Frecuencia del Término** × **Frecuencia Inversa del Documento** conocida como (*Term Frequency - Inverse Document Frequency* ($TF-IDF$)). Para evitar que el valor del término t_{ij} sea constante $\forall d_j \in C$, donde d_j representa un documento y C el conjunto de documentos, (Gerard, 1989) propuso combinar la función $TF(\overrightarrow{t_i}, \overrightarrow{d_j})$ con el factor $IDF(\overrightarrow{t_i}, \overrightarrow{d_j})$ según la ecuación (52), donde f_{ij} denota la frecuencia del rasgo t_i en d_j y se corrige el factor $IDF(\overrightarrow{t_i}, \overrightarrow{d_j})$ de forma que el valor que toma un mismo rasgo en dos documentos, es diferente siempre que la frecuencia de dicho rasgo en cada documento sea también diferente. N representa el número total de documentos del corpus, $df(\overrightarrow{t_i})$ es el número de documentos donde el término aparece (*ej.* $TF(\overrightarrow{t_i}, \overrightarrow{d_j}) \neq 0$), por lo que en caso que el término no exista en ningún documento se asigna valor uno para evitar la división por cero.

$$TF - IDF(\overrightarrow{t_i}, \overrightarrow{d_j}) = \mathrm{TF}(\overrightarrow{t_i}, \overrightarrow{d_j}) \times \log\left(\frac{N}{df(\overrightarrow{t_i})}\right) \quad (52)$$

Entonces, para adaptar esta medida de normalización a los vectores de conceptos relevantes, se toma a cada término como concepto y los documentos corresponden a cada texto de opinión. Esto provoca que $TF(\overrightarrow{t_i}, \overrightarrow{d_j}) =$ valorAR del concepto t_i *y* d_j es el j-ésimo texto (opinión) del corpus de análisis. La función $df(\overrightarrow{t_i})$ se corresponde como el número de opiniones donde aparece el concepto t_i. Entonces, el valor que toma un mismo concepto en las opiniones es diferente, siempre que AR de dicho concepto en cada opinión sea también diferente. Como resultado, son penalizados los conceptos que resulten frecuentes en la colección de textos (C).

5.3.3. APLICACIÓN DE TÉCNICAS DE CLASIFICACIÓN

En este punto del método se han extraído los Conceptos Semánticos Relevantes utilizando las fuentes de WND, SUMO, WNA y WN. Estos se han normalizado, filtrando posibles ruidos que pudieran introducir conceptos muy comunes (ej. dominio *Factotum*). El siguiente paso, es medir el grado de similitud entre cada uno de los vectores característicos de las opiniones, con cada uno de los vectores característicos de las categorías. Obviamente, aquel vector de categoría que ofrezca mayor similitud con el vector del documento (opinión) es el que se asigna a la opinión. En el método MSF-TC se ofrecen dos variantes excluyentes, una es aplicando el algoritmo de K-Medias (*K-Means*) (MacKay, 2003) y la otra es midiendo similitudes de diversas formas y decidiendo la pertenencia de las opiniones a las categorías que maximicen el grado de similitud. A continuación primero se expone la propuesta con K-Media y luego las de asignación directa.

5.3.3.1. APLICACIÓN DE K-MEDIAS

K-Medias es un método de análisis de grupos que tiene como objetivo la partición de n objetos en k grupos en los que cada objeto pertenece al grupo de media más cercana. Partiendo de un conjunto de objetos a clasificar $x_1, x_2 \dots x_p$, el algoritmo realiza las siguientes operaciones en orden lineal:

1. Estableciendo previamente el número de clases existentes (k), se escogen al azar entre los elementos a agrupar de forma que constituyan los centroides de las k clases. Es decir:

$$\alpha_1 : Z_1(1); \alpha_2 : Z_2(1); \dots \ \alpha_k : Z_k(1) \tag{53}$$

El valor dentro del paréntesis corresponde a cada iteración del algoritmo (Véase la Figura 43).

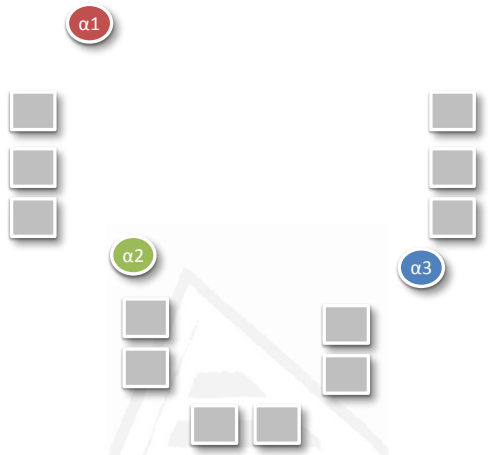


Figura 43. Centroides y Objetos a clasificar con K-Medias.

2. Al tratarse de un proceso recursivo con un contador n, en la iteración genérica n se distribuyen todas las muestras $\{x\}1 \leq j \leq p$ entre las k clases, de acuerdo con la regla correspondiente a la ecuación (54), donde un objeto x pertenece al conjunto de un centroide cuando este tenga con respecto a ese centroide, la menor distancia en comparación con los demás centroides. Véase la Figura 44.

$$x \in \alpha_j(n) \quad si \quad \|x - Z_j(n)\| < \|x - Z_i(n)\| \qquad \forall i = 1, 2 \cdots k \ / \ i \neq j \tag{54}$$

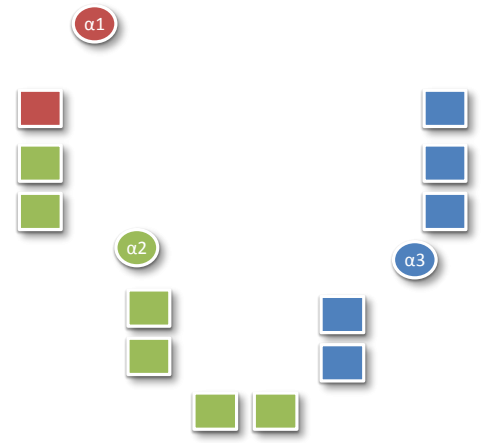


Figura 44. Primera iteración de K-Medias

3. Una vez redistribuidos los elementos a agrupar entre las diferentes clases, es preciso actualizar los centroides de las clases. El objetivo en el cálculo de los nuevos centroides es minimizar el índice de rendimiento siguiente:

$$J_i = \sum_{x \in \alpha_i(n)} \|x - z_i(n)\|^2;\ i\ =\ 1, 2 \dots k \tag{55}$$

Este índice se minimiza utilizando la media muestral o aritmética de α_i: $Z_i(n)$:

$$z_i(n+1) = \frac{1}{N_i(n)} \sum_{x \in \alpha_i(n)} x\ ;\ i\ =\ 1, 2 \dots k \tag{56}$$

Siendo N el número de elementos de la clase α_i en la iteración n (Véase la Figura 45).

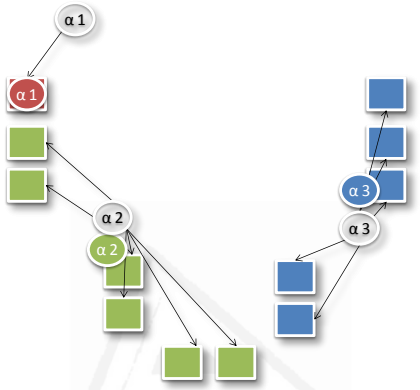


Figura 45. Actualización de los centroides de K-Medias

4. Por último se comprueba si el algoritmo ha alcanzado una posición estable. Si se cumple:

$$z_i(n+1) = z_i(n) \quad \forall\, i = 1, 2 \dots k \tag{57}$$

Si se cumple, el algoritmo finaliza. En el caso contrario vuelve al paso 2 ocurriendo lo nuevamente el mismo proceso tal y como se muestra en la Figura 46.

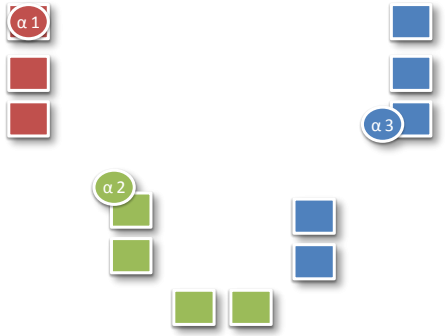


Figura 46. Segunda iteración de K-Medias

Para la aplicación de este algoritmo se considera cada objeto como un documento y cada clase como una categoría. Ambos elementos se representan por vectores de conceptos anteriormente obtenidos con el uso del método RST. Para la actualización del centroide se aplica la ecuación (56) a nivel de conceptos, donde para cada concepto de cada vector se emplea la media muestral. Con el objetivo de asociar las clases a los centroides luego de cada iteración, se necesita utilizar alguna medida de distancia. En correspondencia con este planteamiento, a continuación se describe la distancia Euclidiana utilizada en esta propuesta de K-Medias.

5.3.3.1.1. DISTANCIA EUCLIDIANA

Esta métrica mide la distancia entre dos puntos, donde los vectores de pesos de los términos que representan a los textos son vistos como puntos en el espacio Euclidiano. El valor mínimo entre dos puntos que se obtiene con la distancia Euclidiana es cero y a medida que más se aleje de cero, la distancia entre dos puntos es mayor (Chi and Yan, 1995). Esta métrica es un caso particular de la distancia denominada Minkowski (Mongay, 2005), por lo que hereda sus inconvenientes. Entre ellos, un inconveniente es que influye considerablemente el tamaño de los vectores, considerando también que no se dispone de una diferencia normalizada (que varía entre cero y uno). Para resolver este problema a la distancia Minkowski se le han propuesto diferentes modificaciones a fin de evitar tales inconvenientes. Una de sus variantes es el cálculo de la distancia Euclidiana entre dos puntos y esta se expresa con la Ecuación (58).

$$\mathrm{E}(\vec{x},\vec{y}) = \sqrt[2]{\sum_{\mathrm{i}=1}^{\mathrm{d}}(x_i - y_i)^2} \tag{58}$$

Donde d es el número de términos o coincidencia entre dos vectores, i es el i-ésimo término de cada documento.

Ya conociendo las herramientas para llevar a cabo la clasificación mediante K-Medias se procede a desglosar un ejemplo concreto con el fin de entender su aplicación en MSF-TC. En el siguiente ejemplo se plantean tres vectores de conceptos relevantes teniendo en cuenta que *d* representa a los conceptos de Dominio en este caso en particular. Y su valor asociado es el de *AR* calculado hipotéticamente. Para los vectores con formato [Concepto$_1$, valor AR_1| Concepto$_2$, valor AR_2|...| Concepto$_n$, valor AR_n]:

- $\overrightarrow{N1}$ {d_1, 2|d_2, 3|d_3, 1|d_4, 4|d_5, 6}, corresponde a un centroide
- $\overrightarrow{D2}$ {d_1, 3|d_4, 5|} , corresponde al texto de una opinión
- $\overrightarrow{D3}$ {d_3, 2|d_4, 5|d_5, 7}, corresponde al texto de una opinión

Distancia aplicada a $\overrightarrow{N1}$ y $\overrightarrow{D2}$, en este caso el número de coincidencia es 2, entonces $d = 2$. La ejecución del cálculo se muestra a continuación:

$$E(\overrightarrow{N1},\overrightarrow{D2}) = \sqrt[2]{(d1_{N1} - d1_{D2})^2 + (d4_{N1} - d4_{D2})^2} = \sqrt[2]{\sum_{i=1}^{2}(N1_i - D2_i)^2} = \sqrt[2]{(2-3)^2 + (4-5)^2} \approx 1.4142 \tag{59}$$

Distancia aplicada a $\overrightarrow{N1}$ y $\overrightarrow{D3}$ con *d*=3:

$$\mathrm{E}(\overrightarrow{N1},\overrightarrow{D3}) = \sqrt[2]{\sum_{\mathrm{i}=1}^{3}(N1_i - D3_i)^2} = \sqrt[2]{(1-2)^2 + (4-5)^2 + (6-7)^2} \approx 1.7320 \tag{60}$$

El vector $\overrightarrow{D3}$ tiene mayor coincidencia con $\overrightarrow{N1}$ que $\overrightarrow{D2}$, por lo que $\overrightarrow{D3}$ tiene que ser más cercano a $\overrightarrow{N1}$. La diferencia entre el valor de AR de los dominios es uno, pero la distancia Euclidiana indica lo contrario, favoreciendo más a los vectores que menor coincidencia tengan (sea de recordar que mientras más cerca de cero esté el resultado existe menor distancia). Una posible solución puede ser utilizar una modificación que respecta a la siguiente ecuación (61), donde *p*

para corresponder con la distancia Euclidiana es igual a dos y v es el número de conceptos (términos) que coinciden en los vectores. Con la introducción de v se consideran las coocurrencias de etiquetas de dominios entre documentos a comparar, haciendo que a mayores coincidencias de presencia de conceptos, el valor resultante disminuya y por tanto provoca mayor similitud. Esta solución es la que se ha considerado para las experimentaciones.

$$L_p(\vec{x},\vec{y}) = \sqrt[p]{\frac{\sum_{i=1}^{d}(x_i - y_i)^p}{v}} \tag{61}$$

5.3.3.2. CLASIFICACIÓN POR ASIGNACIÓN DIRECTA

En esta fase se distribuyen todos los objetos a agrupar entre las categorías, para luego aplicar varias medidas de similitud. Los objetos donde, la similitud entre sus vectores de características de una categoría es la mayor, se clasifican asociándose a esa categoría. El proceso de clasificación directa se realiza de forma electiva, es decir, solamente se aplica una de las medidas en cada momento de clasificación, lo que significa que se cuenta con cinco formas más de aplicar TC en Minería de Opiniones. A continuación se enumeran cada una de las medidas de similitud proponiéndose lo que se desea de ellas.

1. Obtención del máximo valor de RSD (Resta del área de similitud con el área de desigualdad entre vectores),
2. Obtención del mínimo valor distancia Euclidiana
3. Obtención del máximo valor del Coeficiente de Correlación de Pearson,
4. Obtención del máximo valor de la Función Jaccard,
5. Obtención del máximo valor del Coseno del ángulo

5.3.3.2.1. RESTA DEL ÁREA DE SIMILITUD CON EL ÁREA DE DESIGUALDAD ENTRE VECTORES

Esta propuesta no es una distancia establecida en la literatura, pero bien se puede usar el área bajo la curva que generan los vectores. Debido a que están los conceptos, ordenados por el eje *x* por orden alfabético y ascendente. En caso de no existir ciertos conceptos, estos se colocan con valor cero pero no cuentan como coincidentes. Si se tienen dos vectores $\overrightarrow{v1}$ y $\overrightarrow{v2}$ cuyos valores son $\overrightarrow{v1}$ $\{d_1, 0|d_2, 2|d_3, 3|d_4, 4|d_5, 2|d_6, 4\}$ *y* $\overrightarrow{v2}$ $\{d_1, 1|d_2, 0.5|d_3, 1.5|d_4, 5|d_5, 6|d_6, 2\}$ y sus representaciones en el espacio corresponden con la Figura 47.

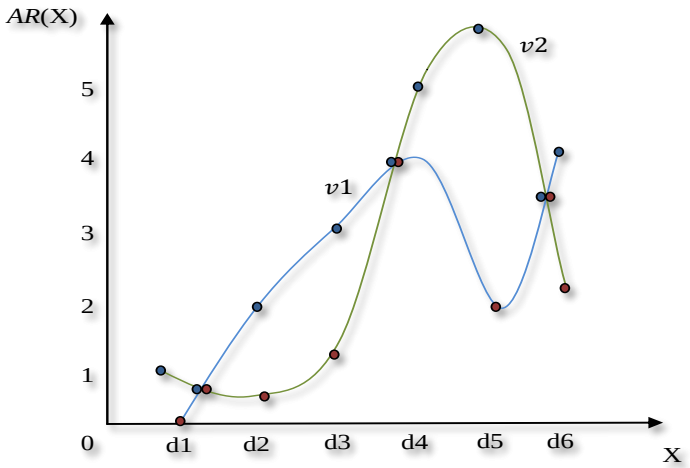


Figura 47. Representación espacial de dos vectores de conceptos y sus valores de AR.

El área comprendida entre los círculos rojos y el eje X, es el área común entre los vectores. Los puntos que delimitan esta área son los mínimos entre los dos vectores y donde estos se intersecan, con el eje X. El área comprendida entre los círculos azules y rojos, es el área de desigualdad entre los vectores, los puntos que delimitan esta área son los máximos entre los dos vectores y donde estos se intersecan, con los puntos mínimos. Es importante señalar que los puntos de intersección (representados por la unión de los puntos rojos y azules), si no se conocen se tienen que buscar para no perder o aumentar el área común.

Para buscar los puntos donde se intersecan los vectores, se toman los dos puntos más cercanos a la intersección de cada vector, de tal manera que cuando se aplique una interpolación, se generen dos polinomios que igualándolos se encuentra el punto deseado. Para aplicar este proceso se siguen los siguientes pasos:

1. Primer paso. Detectar las intersecciones en los vectores

Para detectar las intersecciones se revisan todas las parejas de puntos, por ejemplo para los vectores $\overrightarrow{v1}$ $\{d_1, 0|d_2, 2\}$, $\overrightarrow{v2}$ $\{d_1, 1|d_2, 0.5\}$, d_1 toma un valor menor en $\overrightarrow{v1}$ que en $\overrightarrow{v2}$, pero d_2 toma un valor mayor en $\overrightarrow{v1}$ que en $\overrightarrow{v2}$, entonces entre estos dos puntos existe una intersección.

2. Segundo Paso. Aplicar polinomio de interpolación de Lagrange

En general se plantea que, si se conocen los valores que toma la función $f(x)$ en los $n + 1$ puntos diferentes $x_0, x_1, \dots, x_n$, el problema de interpolación consiste en hallar una función $g(x)$ cuyos valores pueden ser calculados para cualquier x en un intervalo $\{x_0, x_1, \dots, x_n\}$ de manera que $g(x_0) = f(x_0)$, $g(x_1) = f(x_1)$,…, $g(x_n) = f(x_n)$. Los números $x_0, x_1, \dots, x_n$ se les nombran puntos (o nodos) de interpolación. Si x no es un nodo de interpolación, el valor obtenido por $g(x)$ se le llama valor interpolado. Como sugerencia se tiene usar polinomios de grado pequeños (Alvarez and all, 1998), para hacer simples las evaluaciones. La ecuación (62) se refiere a la interpolación de Lagrange, donde se corresponde con los $n + 1$ pares ordenados $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$.

$$P_n(x) = \sum_{i=0}^{n} L_i(x) y_i \tag{62}$$

Donde:

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} = \frac{(x - x_0)}{(x_i - x_0)} \frac{(x - x_1)}{(x_i - x_1)} \dots \frac{(x - x_{i-1})}{(x_0 - x_{i-1})} \frac{(x - x_{i+1})}{(x_0 - x_{i+1})} \dots \frac{(x - x_n)}{(x_0 - x_n)} \tag{63}$$

Una vez conocido donde hay una intersección se escogen los dos puntos de cada vector que se encuentran antes y después de la intersección, y por cada una de las parejas de puntos pertenecientes a un vector, se busca su polinomio de la siguiente manera. Para los vectores $\overrightarrow{v1}$ $\{d_1, 0|d_2, 2\}$, $\overrightarrow{v2}$ $\{d_1, 1|d_2, 0.5\}$ cuyas representaciones son:

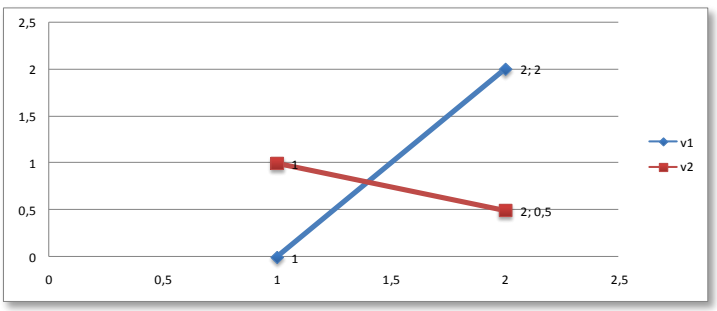


Figura 48. Intersección de polinomios.

El proceso para buscar el polinomio de interpolación para el vector $\overrightarrow{v_1}$ cuyos puntos son (1; 0); (2; 2) es:

- X_0= 1 Y_0=0
- X_1=2 Y_1=2

Para $\overrightarrow{v_2}$ cuyos puntos son (1; 2) y (2; 0,5) es:

- X_0= 1 Y_0=2
- X_1=2 Y_1=0,5

Como tiene dos nodos el grado del polinomio es uno. Luego se buscan los $L_0(x)$, $L_1(x)$ utilizando la ecuación (63) de la siguiente manera:

$$L_0(x) = \frac{(x - x_1)}{(x_0 - x_1)} = \frac{(x - 2)}{(1 - 2)} = -(x - 2) \tag{64}$$

$$L_1(x) = \frac{(x - x_0)}{(x_1 - x_0)} = \frac{(x - 1)}{(2 - 1)} = (x - 1) \tag{65}$$

Acto seguido se obtiene el polinomio con la ecuación (63).

$$P_1(x) = L_0(x)y_0 + L_1(x)y_1 = 2x - 2 \text{ (correspondiente al } \overrightarrow{v_1}) \tag{66}$$

$$P_2(x) = L_0(x)y_0 + L_1(x)y_1 = -x + 2 + 0.5x - 0.5 = -0.5x + 1.5 \text{ (correspondiente al } \overrightarrow{v_2}) \tag{67}$$

3. Tercer paso. Igualar los polinomios

Una vez obtenido los polinomios se igualan y se despeja la variable x. Quedando de la siguiente forma:

$$2x - 2 = -0.5x + 1.5 \tag{68}$$

$$x = 3.5/2.5 = 1.4 \tag{69}$$

4. Cuarto Paso. Obtener el valor de *y*

Para buscar el valor de *AR*(*x*) se evalúa el valor obtenido de *x* en la función del polinomio, quedando de la siguiente forma:

$$P_2(1.4) = -0.5 * 1.4 + 1.5 = 0.8 \tag{70}$$

Una vez obtenido los valores del punto de intersección para $(x; y)$ es posible calcular el área bajo la curva comprendida entre los puntos máximos y el eje *x*, además del área bajo la curva comprendida entre los puntos mínimos y el eje x. Para calcular estas áreas es preciso tener en cuenta el principio del **método de los trapecios**, el cual aplicando la integral definida bajo la curva consigue calcular el área total aproximada, que es la suma de las áreas de los n pequeños trapecios generados bajo el polinomio creado de todos ellos con una de anchura h (la anchura se corresponde con la altura de la ecuación de área de un trapecio). Esta altura consiste en el paso generado según la cantidad de trapecios en los que se desee seccionar las áreas del polinomio (Alvarez and et.al., 2004). Como en este caso se conocen todos los puntos, no es necesario aplicar dicha integral definida y por consiguiente se procede con el cálculo del área de todos los trapecios generados, incluyendo los que se crean a partir de las intercepciones. La ecuación (71) que define el cálculo del área del trapecio.

$$At = \frac{(a+c)}{2}h \tag{71}$$

Donde: *a* corresponde con la base1, *c* con la base2 y h con la altura.

Si se tienen como bases los valores de AR mínimos ($ARmin$ =$\{AR(x_1), AR(x_2), \ldots, AR(x_n)\}$) (para el eje y) incluyendo los que respectan a la intercepción. Los valores correspondientes para el eje x que se nombran ahora $Cmin$ = $\{x_1, x_2, \ldots x_n\}$, donde cada concepto se enumera tomando para el eje x valores consecutivos iniciando en uno e incrementándose con paso uno. Los únicos valores del eje *x* que pueden tener valores distintos de paso uno, son los calculados luego de existir una intercepción entre los polinomios. El cálculo del área total se aplica con la ecuación (72) que se rige por el cálculo de área del trapecio.

$$Amin = \sum_{i=1}^{n-1}\left(\frac{ARmin_{i+1} + ARmin_i}{2} * (Cmin_{i+1} - Cmin_i)\right) \tag{72}$$

Donde *n* representa la cantidad trapecios generados por el polinomio de mínimos valores.

Para comprender mejor la propuesta, a continuación se procesan los vectores tomados anteriormente como ejemplos ($\overrightarrow{v1}\{d_1, 0|d_2, 2\}$, $\overrightarrow{v2}\{d_1, 1|d_2, 0.5\}$).y utilizando la interpolación de que se obtuvo de ellos. Según ilustra la Figura 48 y correspondiendo con el ejemplo desarrollado de intercepción, se tienen como mínimos los puntos (1, 0), (1.4, 0.8), (2; 0.5) conformando las siguientes colecciones $ARmin$ = $\{0, 0.8, 0.5\}$ y $Cmin$ = $\{1, 1.4, 2\}$ siendo (1.4, 0.8) un punto de intercepción. Entonces el área común entre los dos vectores, se encuentra entre estos puntos y el eje x. Para calcular el área bajo la curva de los mínimos y los máximos valores la ecuación (73) ofrece el siguiente resultado:

$$\begin{aligned} Amin &= \left(\frac{0.8+0}{2} * (1.4-1)\right) + \left(\frac{0.5+0.8}{2} * (2-1.4)\right) \\ &= \left(\frac{0.8*0.4}{2}\right) + \left(\frac{1.3*0.6}{2}\right) = 0.16 + 0.39 = 0.55u^2 \end{aligned} \tag{73}$$

Los máximos en este caso son $ARmax$ =$\{1, 0.8, 2\}$ siendo 0.8 una intersección, entonces el área entre estos puntos y el eje x es:

$$\begin{aligned} Amax &= \left(\frac{0.8+1}{2} * (1.4-1)\right) + \left(\frac{2+0.8}{2} * (2-1.4)\right) \\ &= \left(\frac{1.8*0.4}{2}\right) + \left(\frac{2.8*0.6}{2}\right) = 0.36 + 0.84 = 1.2u^2 \end{aligned} \tag{74}$$

Una vez obtenidas estas áreas se puede adquirir el área de desigualdad entre los vectores, restándole a A_{max} , A_{min} .

$$A_{desigualdad} = A_{max} - A_{min} \tag{75}$$

El área de igualdad entre los vectores es A_{min} , esta indica el grado de similitud. Pero no, cuánto no se parecen, es por eso que se necesita el área de desigualdad. El valor real de igualdad entre estos vectores es el área de igualdad menos el de desigualdad, ya que pueden existir dos vectores que tengan un valor de igualdad considerable pero también pueden ser muy desiguales. Todo esto se resume con la expresión.

$$ValorCalc = A_{min} - A_{desigualdad} \tag{76}$$

5.3.3.2.2. OBTENCIÓN DEL MÍNIMO VALOR DISTANCIA EUCLIDIANA

Al aplicar la clasificación mediante el uso de la distancia Euclidiana solamente, se aplica el proceso comentado en la sección 5.3.3.1.1. Le será asignada la categoría al documento que minimice esta distancia al comparar vectores de conceptos.

5.3.3.2.3. OBTENCIÓN DEL MÁXIMO VALOR DEL COEFICIENTE DE CORRELACIÓN DE PEARSON

El coeficiente de correlación de Pearson se define como un índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas. El coeficiente se establece en el rango $\{-1 \ldots +1\}$, indicando un grado de fortaleza tanto para valores próximos a $+1$ como a -1. Al asociarse al primero se indica que una relación es perfecta positiva y en el segundo que es perfecta negativa. Por esta razón, la correlación entre dos variables es perfecta positiva en la medida exacta en que aumenta una y aumenta la otra. Se dice entonces que es perfecta negativa a medida que una aumenta exactamente la otra disminuye (Achen, 2008). La ecuación (77) pone en práctica esta medida.

$$C_{pearson} = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\left[\sum_{i=1}^{n}(x_i - m_x)^2\right]\left[\sum_{i=1}^{n}(y_i - m_y)^2\right]}} \tag{77}$$

Donde:

$$m_x = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \tag{78}$$

$$m_y = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) \tag{79}$$

Las variables x_i y y_i representan las relevancias de una característica determinada en los documentos x, y, donde n es la dimensión de los vectores de representación.

El valor del coeficiente de correlación entre dos variables se puede transformar si se desea en valores de correlación en el intervalo $\{0 \ldots 1\}$ mediante la transformación:

$$C^*_{pearson} = \frac{(1 + C_{pearson})}{2} \tag{80}$$

5.3.3.2.4. OBTENCIÓN DEL MÁXIMO VALOR DE LA FUNCIÓN JACCARD

Conocido como función de similitud de Jaccard, esta constituye una función estadística que compara los términos que hay en común entre dos vectores que representan a dos documentos (Chen *et al.*, 1998). Si dos documentos al ser representados en el espacio vectorial por un vector *t*-dimensional, muestran términos en común, la representación de estos documentos es semejante y se estima que los documentos lo son también. Al comparar dos vectores se obtiene un número entre cero y uno. Cuando se obtiene un valor resultante igual a uno indica que los vectores son completamente semejantes y por el contrario con cero son completamente

diferentes. En la ecuación (81) $sim\ (\vec{x}, \vec{y})$ representa la medida de semejanza, donde $\vec{x}$ y $\vec{y}$ son vectores representativos de los documentos a comparar, *n* es el número de coincidencias entre dos vectores y k es el k-ésimo término de cada vector.

$$sim(\vec{x}, \vec{y}) = \frac{\sum_{k=1}^{n}(x_k . y_k)}{\sum_{k=1}^{n}({x_k}^2) + \sum_{k=1}^{n}({y_k}^2) - \sum_{k=1}^{n}(x_k . y_k)} \quad (81)$$

Al aplicar esta medida entre los vectores correspondientes a documentos y categorías se asocian los que maximicen este valor.

5.3.3.2.5. OBTENCIÓN DEL MÁXIMO VALOR DEL COSENO DEL ÁNGULO

El uso de esta medida presentada por la ecuación (4) y de tal manera que las demás medidas implicadas en la clasificación, se obtienen valores capaces de representar vectorialmente un espacio de amplitud. Donde, a mayor valor obtenido entre vectores, mayor es la similitud. Consecuentemente se asigna al documento, la categoría que maximice este valor.

5.4. EVALUACIONES Y RESULTADOS

Las evaluaciones de la aplicación del Análisis Semántico Multidimensional en tareas de opiniones se aplican sobre los corpus de las tareas del inglés monolingüe de la competición del MOAT. En esta competición a los participantes se les proporcionó veinte temas. Para cada uno de los temas, se asignaron una serie de documentos divididos en oraciones para la aplicación de las tareas de relevancia, existencia de opinión y determinación de polaridad. En particular, para la evaluación de las propuestas presentadas en este capítulo, cada frase contenida en un documento se toma como documento en el proceso de clasificación de opiniones. Es decir cada vez que se use el término documento se hace referencia a frase. A continuación se describen dos tipos de evaluaciones:

- La primera consiste en evaluar la propuesta descrita en la sección 5.2 con respecto a detección a la detección de opiniones, relevancia y polaridad.
- En la segunda se evalúan un conjunto de aproximaciones que basadas en el Análisis Semántico Multidimensional que proponen la clasificación de las opiniones como frases respecto a cada uno de los tópicos del corpus de la competición MOAT.

5.4.1. EVALUACIÓN DEL ANÁLISIS DE OPINIONES CON RESPECTO A LA DETECCIÓN DE OPINIONES, RELEVANCIA Y POLARIDAD

En esta sección las evaluaciones se han concentrado en determinar la influencia de cada dimensión (de recursos) por separado y conjuntamente a favor de la propuesta Senti-RST. Además, se establece una comparación con los mejores resultados obtenidos por los sistemas participantes en la competición de MOAT NTCIR 8.

5.4.1.1. ANÁLISIS DE LA INFLUENCIA DE LAS DIMENSIONES SEMÁNTICAS NORMALIZANDO VECTORES

En esta sección se presentan los resultados de las tres tareas descritas mediante la combinación de todas las dimensiones combinadas y el uso de cada una por separado. Más

adelante, se describen los experimentos mostrando sus respectivos resultados en la Tabla 47. Los cálculos de *Precision* y *Recall* se obtienen a partir del *scorer* que proporciona la competición.

- Exp 1: Combinando todas las dimensiones (WND, WNA, taxonomía de WN, SUMO y SC)
- Exp 2: Usando solamente WNA
- Exp 3: Usando solamente WND
- Exp 4: Usando solamente SC
- Exp 5: Usando solamente SUMO
- Exp 6: Usando solamente la taxonomía de WN

Exp	*Opinion*			*Relevance*			*Polarity*		
	P	R	F	P	R	F	P	R	F
Exp 1	0.206	0.878	0.333	0.788	0.868	**0.826**	0.394	0.345	**0.368**
Exp 2	0.238	0.572	0.336	0.779	0.558	0.651	0.397	0.222	0.285
Exp 3	0.226	0.695	**0.341**	0.794	0.692	0.740	0.403	0.275	0.327
Exp 4	0.201	0.885	0.333	0.788	0.873	**0.828**	0.397	0.349	**0.372**
Exp 5	0.213	0.865	**0.342**	0.790	0.858	**0.823**	0.406	0.337	**0.368**
Exp 6	0.211	0.876	**0.341**	0.788	0.866	**0.825**	0.405	0.342	**0.371**

Tabla 47. Resultados con el uso de vectores normalizados. *Precision* (*P*)*, Recall* (*R*) *y F-Measure* (*F*).

Los mejores resultados se obtienen en el experimento 4 y 6, correspondientes a la taxonomía de WN y SC. Sin embargo, los otros resultados son similares en el nivel de rendimiento. Esto indica que la propuesta puede ser aplicada con éxito en tareas de Minería de Opinión sin problema alguno, al incluir diferentes dimensiones semánticas.

5.4.1.2. INFLUENCIA DE LAS DIMENSIONES SEMÁNTICAS SIN NORMALIZAR VECTORES

Con el fin de demostrar que la normalización de vectores afecta los resultados, se presentan los mismos experimentos sin haber aplicado el proceso de normalización. La Tabla 48 muestra los resultados:

- Exp 7: Combinando todas las dimensiones (WND, WNA, taxonomía de WN, SUMO y SC)
- Exp 8: Usando solamente WNA
- Exp 9: Usando solamente WND
- Exp 10: Usando solamente SC
- Exp 11: Usando solamente SUMO
- Exp 12: Usando solamente la taxonomía de WN

Exp	*Opinion*			*Relevance*			*Polarity*		
	P	R	F	P	R	F	P	R	F
Exp 7	0.201	0.885	0.333	0.788	0.873	**0.828**	0.397	0.349	0.372
Exp 8	0.233	0.611	0.337	0.784	0.600	0.680	0.423	0.255	0.318
Exp 9	0.219	0.779	**0.342**	0.792	0.773	0.782	0.394	0.305	0.344
Exp 10	0.206	0.877	0.334	0.789	0.867	**0.826**	0.446	0.389	**0.416**
Exp 11	0.206	0.850	0.332	0.785	0.836	**0.810**	0.446	0.377	**0.409**
Exp 12	0.205	0.855	0.331	0.787	0.844	**0.815**	0.437	0.370	**0.401**

Tabla 48. Resultados para cada tarea sin normalizar vectores. *Precision* (*P*)*, Recall* (*R*) *y F-Measure* (*F*).

En la Tabla 48, se muestra los valores de F-Medida obtenidos para polaridades, donde todos mejoran a los resultados anteriores que normalizaban los vectores. Es importante anotar que no ayuda la normalización de los vectores de la tarea de clasificación de polaridad. Todos los experimentos presentados en la Tabla 48 mejoran los anteriores y la experimentación con SC obtiene los mejores resultados en polaridad y en relevancia. Es importante resaltar que los

comportamientos de todas las dimensiones están equilibrados, incluso cuando se utilizan en conjunto. Esto indica que para el análisis de opiniones es válido el uso de cualquiera de estos recursos.

5.4.1.3. COMPARACIÓN CON OTRAS PROPUESTAS

En esta sección se presenta una comparación entre Senti-RST y los mejores sistemas que participaron en NTCIR 8 MOAT. Comenzando con el análisis de la tarea de detección de opinión en las frases, los únicos sistemas que obtienen superiores resultados en comparación con Senti-RST son los sistemas UNINE (Zubaryeva and Savoy, 2010) y NECLC (véase la Tabla 49). Estos ostentan *F-valores* de 40,1% y 36,52% respectivamente. Con una diferencia respecto al mejor de los experimentos de 5.9% y 2.32% respectivamente.

Para conocer las peculiaridades de UNINE y poder marcar puntos de convergencia y divergencia se analiza su propuesta. Esta se basa en la selección de los términos relevantes que claramente pertenecen a un tipo de polaridad. Para obtener primeramente estos términos relevantes, UNINE calcula la probabilidad de la ocurrencia de la palabra *w* en todo el corpus *C*. Luego, computa un $Z\_scrore\,(w)$ y mide el grado de importancia del término. Si $Z\_scrore(w) > número\,heurístico$ se selecciona el término para luego definir las polaridades asociadas a este término, sumando el número valores de polaridades que presenta el término cada vez que este aparece en la frase. La puntuación de opiniones final se consigue con la suma de resultados positivos y negativos para cada término seleccionado. El valor de puntuaciones para determinar frases que no contienen opinión, se computa con la suma de las puntuaciones de objetividad para cada término seleccionado, dividido por el número de palabras en la frase. Estos tres parámetros (puntuaciones positivas, negativas y objetivas) se introducen en un sistema de entrenamiento para asignar la condición correcta.

Senti-RST no tiene en cuenta la detección de los términos (palabras) relevantes, ni las puntuaciones objetivas. UNINE también obtiene mejores resultados que Senti-RST en la tarea de polaridad, surgiendo la idea que la combinación de esta propuesta con Senti-RST podría ofrecer importantes resultados. Teniendo en consideración que ambos aplicaron la extracción de características relevantes, UNINE extrayendo características léxicas y Senti-RST características semánticas.

En la tarea de polaridad, Senti-RST ofrece resultados similares a los de la primera ejecución del sistema de UNINE, con aproximadamente un 37% de F-Medida. Pero sin embargo con cierto margen de diferencia de las mejores salidas del mismo sistema (un 51,03% de F-Medida) (véase la Tabla 49). Para la tarea de relevancia, Senti-RST acierta con una diferencia de F-Medida = 3,22% respecto a los mejores resultados de todas las ejecuciones presentadas por la Universidad Nacional de Taiwan (NTU). Siendo este, el mejor de todos y demostrando que la cobertura asociada a Senti-RST se beneficia de la información semántica obtenida entre la pregunta y la frase analizada para esta tarea. Senti-RST brinda resultados que podrían colocarlo entre los primeros lugares entre las cuatro tareas bajo la premisa de multidimensionalidad semántica.

Team_ID	***Opinion***			***Relevance***			***Polarity***		
	P	R	F	P	R	F	P	R	F
NTU (1)	0.168	0.955	0.286	0.780	0.961	0.861	0.522	0.499	0.510
NTU (2)	0.170	0.937	0.288	0.778	0.940	0.851	0.492	0.462	0.477
UNINE (2)	0.193	0.818	0.313	0.844	0.360	0.505	0.484	0.378	0.424
Exp 10	0.206	0.877	0.334	0.789	0.867	0.826	**0.446**	**0.389**	**0.416**
Exp 11	0.206	0.850	0.332	0.785	0.836	0.810	**0.446**	**0.377**	**0.409**
Exp 12	0.205	0.855	0.331	0.787	0.844	0.815	**0.437**	**0.370**	**0.401**
UNINE (1)	0.294	0.628	0.401	0.837	0.327	0.471	0.503	0.296	0.373
Exp 4	0.201	0.885	0.333	0.788	0.873	0.828	0.397	0.349	0.372
Exp 7	0.201	0.885	0.333	0.788	0.873	0.828	0.397	0.349	0.372
Exp 6	0.211	0.876	0.341	0.788	0.866	0.825	0.405	0.342	0.371

Exp 5	0.213	0.865	0.342	0.790	0.858	0.823	0.406	0.337	0.368
Exp 1	0.206	0.878	0.333	0.788	0.868	0.826	0.394	0.345	0.368
Exp 9	0.219	0.779	0.342	0.792	0.773	0.782	0.394	0.305	0.344
Exp 3	0.226	0.695	0.341	0.794	0.692	0.740	0.403	0.275	0.327
Exp 8	0.233	0.611	0.337	0.784	0.600	0.680	0.423	0.255	0.318
Exp 2	0.238	0.572	0.336	0.779	0.558	0.651	0.397	0.222	0.285
OPAL (2)	0.194	0.440	0.270	0.826	0.516	0.971	0.509	0.123	0.198
OPAL (1)	0.180	0.452	0.257	0.821	0.478	0.604	0.381	0.128	0.192
NECLC (bsf)	0.265	0.587	0.365						
NECLC (bs0)	0.259	0.581	0.358						
NECLC (bs1)	0.218	0.788	0.341						
KLELAB (3)	0.197	0.680	0.305						
KAIST (2)	0.190	0.653	0.294						
KLELAB (2)	0.179	0.820	0.294						
KAIST (1)	0.189	0.649	0.292						
KLELAB (1)	0.168	0.954	0.286						
OPAL (3)	0.194	0.440	0.270	0.763	0.394	0.749			
PolyU (1)	0.246	0.215	0.229						
SICS (1)	0.139	0.314	0.192						
PolyU (2)	0.195	0.134	0.159						

Tabla 49. Resultados de MOAT ordenados por *F-Measure* de la Polaridad. *Precision* (*P*)*, Recall* (*R*) *y F-Measure* (*F*).

5.4.2. EVALUACIÓN DE LA CLASIFICACIÓN DE TEXTOS APLICADO A OPINIONES

Para evaluar las propuestas de clasificación de textos cortos. Se ha escogido el corpus de opiniones que proporciona la competición del MOAT. En este caso se considera que un texto se encuentra enmarcado dentro de una categoría, debido a que según MOAT las opiniones se organizan dentro de tópicos. De este corpus se escogen diez tópicos con un promedio de 142 opiniones por cada una, en total 1422 opiniones. Estas opiniones son textos muy cortos conteniendo como promedio aproximado doce palabras, lo cual no es favorable para la clasificación tradicional, por ser poco descriptivas de la semántica del texto. Se toma como texto que describe una categoría, la pregunta del tema y las frases que se encuentran dentro de este deben ser clasificadas con relación a la pregunta. Primeramente se separan todos los textos de todos los tópicos seleccionados para la evaluación y se someten al proceso de clasificación midiendo exactitudes según las ecuaciones (82), (83) y (29). A continuación se muestra un ejemplo para que se tenga una idea de la longitud a la que se enfrenta el proceso de clasificación.

- **Categoría (N04):**"*What reasons have been given for the Space Shuttle Columbia accident in February*, 2002?"
- **Texto:** "*If they had the tools, the trend analysis and the engineering analysis, they might have stood up in these meetings and offered checks and balances.*"

5.4.2.1. EVALUACIONES

Para evaluar los resultados se utilizan las funciones de evaluación Precisión, Cobertura, F-Medida, correspondientes a las ecuaciones (82), (83) y (29) respectivamente. Las definiciones de las medidas requieren de los siguientes elementos (véase Tabla 50):

- Verdaderos Positivos: cuando el sistema clasifica un documento como perteneciente a una clase a la que sí pertenece. Cantidad a.
- Verdaderos Negativos: cuando el sistema clasifica un documento como perteneciente a una clase a la que no pertenece. Cantidad b.

- Falsos Positivos: cuando el sistema no clasifica un documento como perteneciente a una clase a la que sí pertenece. Cantidad c.
- Falsos Negativos: cuando el sistema no clasifica un documento como perteneciente a una clase a la que no pertenece. Cantidad d.

	Datos Positivos	Datos Negativos
Asignaciones Positivas	a	c
Asignaciones Negativas	b	d

Tabla 50. Tabla de contingencia.

A partir de la tabla de contingencia se pueden definir las siguientes funciones de evaluación para todo sistema de clasificación:

$$Precision, P_k = \frac{a_k}{(a_k + b_k)} \tag{82}$$

Dada una clase k, la precisión (*Precision*) representa la fracción de asignaciones correctas frente al total de asignaciones positivas realizadas para esa clase.

$$Recall, R_k = \frac{a_k}{(a_k + c_k)} \tag{83}$$

La cobertura (*Recall*) representa la fracción de asignaciones positivas respecto al conjunto real de elementos pertenecientes a la clase k que se esté considerando.

5.4.2.2. CATEGORIZACIÓN CON K-MEDIAS COMBINADA CON DISTANCIA EUCLIDIANA

Para decidir qué documento corresponde a cada categoría se utiliza el proceso definido en la sección 5.3.3.1 (K-Medias y la distancia Euclidiana), con los cuatro vectores correspondientes a las cuatro jerarquías (WN, WND, WNA, SUMO). Con esta prueba se quiere verificar cuál es la combinación de recursos que ofrece más aciertos. También se desea comprobar si la aplicación de la función de normalización ($TF - IDF$) favorece los resultados.

Categorías y breve descripción	R	P	F	Documentos
N04 (*Columbia Accident*)	0.67	0.74	0.71	136
N05 (*Iraq war and Baghdad invasion*)	0.58	0.55	0.56	181
N06 (*SARS Outbreak*)	0.71	0.61	0.66	121
N08(*Madrid, Spain train terrorist attack*)	0.11	0.19	0.14	161
N11(*George W. Bush Presidential Election*)	0.37	0.28	0.32	151
N16(*Anti-Japanese demonstrations in China*)	0.51	0.26	0.35	201
N18(*Hurican Katrina*)	0.34	1	0.51	156
N27(*Direct flights between China and Taiwan*)	0.54	0.60	0.57	93
N32(*diet weight*)	0.52	1	0.68	121
N41(*U.S. military abuse of Iraqi prisoners*)	0.57	0.56	0.56	101
Promedio	**0.49**	**0.58**	**0.50**	**1422**

Tabla 51. Resultados utilizando la combinación de WND y SUMO sin normalización.

En esta prueba se puede observar que los resultados no superan el 50% como promedio en la F-Medida. Estos se encuentran sujetos en gran medida a la escasa información que aporta el contexto de los documentos, ya que, el número de palabras proporcionadas por el documento no son suficientes para obtener una buena información contextual. Como nota importante se tiene que no se aplica la normalización.

Los mismos experimentos de la Tabla 52 se aplican utilizando una función de normalización sobre los vectores de conceptos, y solamente con los conceptos de WND y SUMO.

Categorías y breve descripción	*Recall*	*Precision*	*F- Measure*	Documentos
N04 (*Columbia Accident*)	0.14	0.25	0.18	136
N05 (*Iraq war and Baghdad invasion*)	0.24	0.37	0.29	181
N06 (*SARS Outbreak*)	0.25	0.25	0.25	121
N08(*Madrid, Spain train terrorist attack*)	0.09	0.13	0.11	161
N11(*George W. Bush Presidential Election*)	0.01	0.04	0.02	151
N16(*Anti Japanese demonstrations in China*)	0.05	0.11	0.07	201
N18(*Hurican Katrina*)	0.33	0.77	0.47	156
N27(*Direct flights between China and Taiwan*)	0.67	0.12	0.21	93
N32(*diet weight*)	0.40	0.96	0.57	121
N41(U.S. *military abuse of Iraqi prisoners*)	0.07	0.05	0.06	101
Promedio	**0.22**	**0.30**	**0.22**	**1422**

Tabla 52. Resultados utilizando la combinación de todos los vectores, con la intervención de función de normalización.

Según la Tabla 52 solamente se obtuvo un 22% no superando a la prueba anterior en la F-Medida. Esto demuestra que la necesidad de reducir importancia a conceptos muy frecuentes en los vectores, provoca la pérdida de información vital. Por este motivo de ahora en adelante ninguna experimentación se somete a la normalización $TF - IDF$.

Para conocer la influencia de las dimensiones, se aplican evaluaciones donde se examinan por separado a WND y a SUMO. Según la Tabla 53, estas por independientes no son capaces de superar el 18% y 31% respectivamente, su poder está cuando actúan juntas (véase la Figura 49).

Combinación de Vectores	*Recall*	*Precision*	*F- Measure*
WND, SUMO	**0.49**	**0.58**	**0.5**
Todos (WN, WND, WNA, SUMO)	0.22	0.3	0.22
SUMO	0.3	0.4	0.31
WND	0.16	0.25	0.18

Tabla 53. Resultados de Clasificación con K-Medias midiendo influencia de recursos sin vectores normalizados.

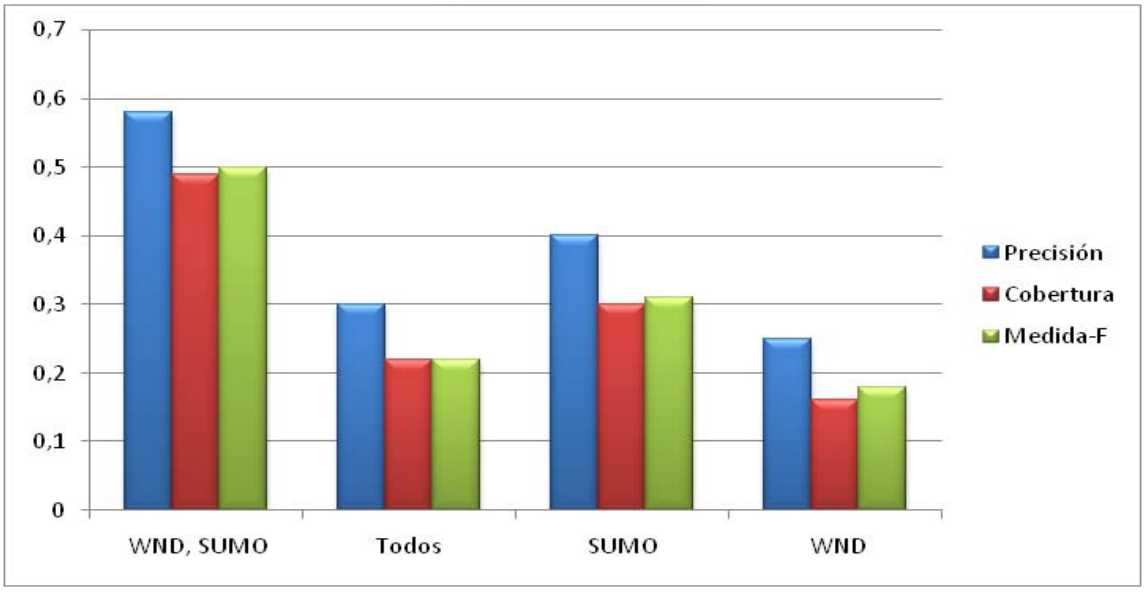


Figura 49. Resultados generales de K-Medias.

Con estas pruebas se demuestra que la utilización de los conceptos relevantes, a partir del uso de WND, SUMO, WNA y WN para definir vectores de características, queda reducido a la utilización de WND y SUMO. Además de eliminar del proceso general el paso intermedio concerniente a la normalización $TF - IDF$.

5.4.2.3. CLASIFICACIÓN POR ASIGNACIÓN DIRECTA.

Tras haber obtenido ciertas conclusiones respecto a qué dimensiones actúan con más fortaleza para la similitud semántica. Se realizan nuevas evaluaciones con relación a los procesos de asignación directa descritos en la sección 5.3.3.2. Solamente se utiliza la combinación de WND y SUMO sin la normalización de vectores. Con estos experimentos se desea verificar, cuáles de las medidas antes mencionadas brindan mejores resultados a la hora de comparar los vectores de conceptos. A continuación se muestran las Tabla 54 y Tabla 55 correspondientes al uso de la distancia Euclidiana y Jaccard respectivamente.

Categorías y breve descripción	*Recall*	*Precision*	*F-Measure*	Documentos
N04 (*Columbia Accident*)	0.18	0.73	0.29	136
N05 (*Iraq war and Baghdad invasion*)	0.34	0.17	0.23	181
N06 (*SARS Outbreak*)	0.28	0.45	0.35	121
N08(*Madrid, Spain train terrorist attack*)	0.17	0.25	0.20	161
N11(*George W. Bush Presidential Election*)	0.07	0.14	0.10	151
N16(*Anti Japanese demonstrations in China*)	0.36	0.20	0.26	201
N18(*Hurican Katrina*)	0.13	0.59	0.21	156
N27(*Direct flights between China and Taiwan*)	0.50	0.31	0.39	93
N32(*diet weight*)	0.47	0.96	0.63	121
N41(U.S. *military abuse of Iraqi prisoners*)	0.41	0.28	0.33	101
Promedio	**0.29**	**0.40**	**0.29**	**1422**

Tabla 54. Resultados de Clasificación utilizando distancia Euclidiana.

Categorías y breve descripción	*Recall*	*Precision*	*F-Measure*	Documentos
N04 (*Columbia Accident*)	0.22	0.55	0.32	136
N05 (*Iraq war and Baghdad invasion*)	0.41	0.24	0.30	181
N06 (*SARS Outbreak*)	0.31	0.40	0.35	121
N08(*Madrid, Spain train terrorist attack*)	0.20	0.24	0.22	161
N11(*George W. Bush Presidential Election*)	0.14	0.14	0.14	151
N16(*Anti Japanese demonstrations in China*)	0.03	0.12	0.05	201
N18(*Hurican Katrina*)	0.25	0.47	0.33	156
N27(*Direct flights between China and Taiwan*)	0.66	0.20	0.31	93
N32(*diet weight*)	0.52	0.65	0.57	121
N41(U.S. *military abuse of Iraqi prisoners*)	0.05	0.13	0.08	101
Promedio	**0.27**	**0.31**	**0.26**	**1422**

Tabla 55. Resultados de Clasificación utilizando Función Jaccard.

Como se puede apreciar ambas obtienen resultados bajos con respecto al uso de K-Medias, pero en comparación con el Coeficiente de Correlación de Pearson (obtiene un 19% de *F-Measure*), el valor RSD (obtiene un 13% de *F-Measure*) y la función Coseno (obtiene un 20% de *F-Measure*) estos son muy superiores. La Tabla 56 muestra los resultados generales obtenidos por estas evaluaciones asociados a la gráfica de la Figura 50. Donde nuevamente se aprecia que el uso de la distancia Euclidiana, ofrece los mejores resultados en la determinación de similitudes vectoriales, obteniendo una precisión muy por encima que las demás.

Distancias	*Precision*	*Recall*	*F-Measure*
Euclidiana	**0.40**	**0.29**	**0.29**
Jaccard	0.31	0.27	0.26
Coseno	0.23	0.22	0.20
Pearson	0.24	0.21	0.19
RSD	0.21	0.14	0.13

Tabla 56. Resultados comparativos entre métodos de asignación directa.

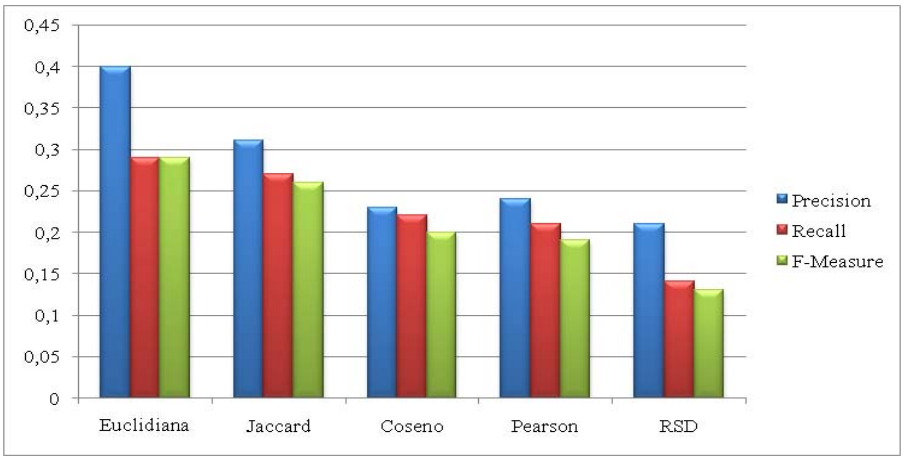


Figura 50. Resultados comparativos entre métodos de asignación directa.

5.4.2.3.1. COMPARACIÓN ENTRE TODOS LOS RESULTADOS DE LOS PROCESOS DE CLASIFICACIÓN

Una vez evaluado los dos procesos generales de clasificación definidos en este capítulo. Solamente falta por establecer una comparativa para conocer cuál es el método más adecuado para clasificar textos cortos. Por una parte se tiene el uso del algoritmo de K-Medias combinado con la distancia Euclidiana donde se toma el mejor resultado obtenido (uso combinado de WND y SUMO sin normalizar vectores). Por otro lado, se toma la mejor variante de asignación directa (distancia Euclidiana uso combinado de WND y SUMO sin normalizar vectores). Al juzgar la gráfica de la Figura 51 los mejores resultados los brinda el proceso que utiliza el algoritmo de agrupamiento, con una diferencia bastante marcada de alrededor de un 10% en *F-Measure*. El valor de precisión ostentado por la mejor propuesta se aproxima al 60% , resultado que todavía se coloca muy lejano a los reportados por investigadores del área de clasificación textual (Li and Sun, 2007, Schneider, 2005, Thorsten, 1998). En este caso se ha de hacer una muy importante observación, investigadores del área aplican sus aproximaciones sobre textos sumamente descriptivos, obteniendo buenos resultados al analizar documentos con vectores de características de más de 500 elementos (Li and Sun, 2007). Además en la mayoría de los casos se aplica aprendizaje automático, por tanto, se necesita un vocabulario muy amplio para que el aprendizaje sea efectivo.

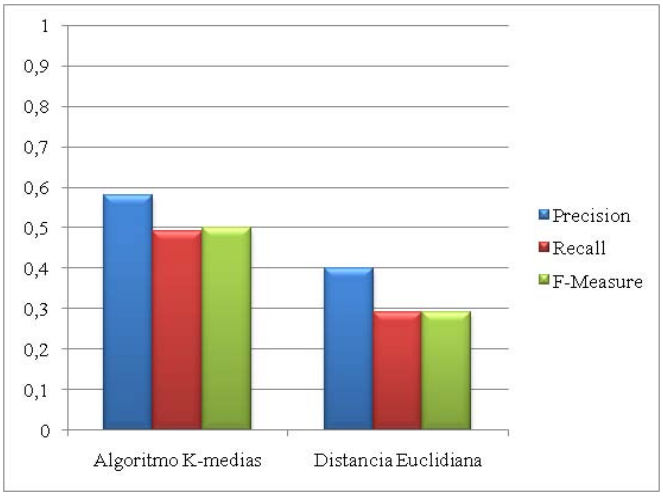


Figura 51. Comparativa entre los mejores resultados de K-Medias y Asignación directa usando WND y SUMO.

5.5. CONCLUSIONES

Motivado por las revelaciones conceptuales que proporciona el hacer uso del Análisis Semántico Multidimensional, en este capítulo se han desarrollado nuevas aplicaciones de la multidimensionalidad semántica sobre áreas de investigación como son la Minería de Opiniones y la Clasificación Textual. Se han analizado estas tareas desde la perspectiva de resolver la clasificación de los textos según polaridades, presencia de opinión y/o pertenencia a un tópico o frase que la describa, utilizando técnicas aplicadas a WSD. Bajo el uso del corpus de la competición MOAT y regido por tres de las tareas que allí se publican, se propone la aproximación Senti-RST. Esta se basa en la construcción de Árboles Semánticos Relevantes capaces de describir textos a nivel conceptual, esta propuesta consigue representar de igual manera Árboles Conceptuales Positivos y Negativos, refiriéndose con ellos a conocer cuánto sentimiento de cada tipo de polaridad alberga el texto analizado. La evaluación se ha aplicado a las tareas de detección de relevancia de un texto de acuerdo con una pregunta, detección de opinión y detección de la polaridad. Donde tras emplear Senti-RST se demuestra que al normalizar los vectores de *AR* se desestima información vital. El comportamiento de la propuesta utilizando cada dimensión semántica aislada y en conjunto, ofrece similares resultados entre sí, haciendo ver que esta técnica es válida para todos aunque las Clases Semánticas ofrecen no por mucho los mejores resultados. Al establecer una comparativa entre Senti-RST y los mejores sistemas que participan en NTCIR 8 MOAT, se destaca que se obtienen resultados que podrían colocarlo entre los primeros lugares entre las tres tareas, introduciendo la multidimensionalidad semántica en el área de análisis de opiniones según muestra la comparativa de la Tabla 49. Algunos sistemas que son capaces de extraer términos relevantes de los textos asociados distintivamente a un tipo de polaridad en particular, superan por muy poco a Senti-RST. Esto indica que una combinación de extracción de términos relevantes y conceptos semánticos relevantes en Senti-RST podría ofrecer importantes resultados.

Por otra parte, también haciendo uso de los textos de la competición MOAT, se proponen varias aproximaciones que se enfrentan a una de las problemáticas concernientes a la TC: la corta longitud del texto a analizar. Tras reconocer de modo general qué características utilizan los sistemas que trabajan en esta área, se proponen nuevas aproximaciones que rompen con los métodos tradicionales. Para ello, se toma como información relevante en un texto los conceptos que RST sea capaz de obtener, y luego se le aplican una serie de medidas de similitud entre vectores de características con el fin de propiciar la TC. Para sus evaluaciones, primeramente se realizaron experimentaciones con varias dimensiones aisladas y en conjunto. Lo que demuestra además que el uso de los recursos WND y SUMO combinados sin establecer normalizaciones del tipo $IF - IFD$, ofrece sus mejores resultados. Esas evaluaciones iniciales se llevan a cabo mediante el uso del algoritmo de K-Medias combinado con la distancia Euclidiana. Después se realizaron nuevas evaluaciones solamente considerando WND y SUMO sin normalización $IF - IFD$, aplicando asignaciones directas de texto al tópico. Entre las medidas implicadas están la distancia Euclidiana, Jaccard, coeficiente de correlación de Pearson, el valor de área RSD (introducida en esta Tesis) y la función Coseno. Los mejores resultados los ofrece el proceso que utiliza el algoritmo de agrupamiento K-Medias (véase la Figura 51). El mejor de los resultados de este se aproxima al 60% , valor que se coloca muy lejano a los reportados por investigadores del área de clasificación textual (Li and Sun, 2007, Schneider, 2005, Thorsten, 1998). Pero se debe aclarar, que los investigadores del área aplican sus aproximaciones sobre textos sumamente descriptivos, obteniendo buenos resultados al analizar documentos con vectores de características de más de 500 elementos (Li and Sun, 2007) y por lo general utilizan sistemas de entrenamiento que necesitan vocabularios muy amplios para que el aprendizaje sea efectivo. MSF-TC se basa en conocimiento, sin requerir de tales informaciones.

6. CONCLUSIONES Y TRABAJOS FUTUROS

Como resultado de esta Tesis doctoral, se presentan las conclusiones generales obtenidas al hacer uso del Análisis Semántico Multidimensional en diferentes tareas del PLN. Se muestran además una serie de propuestas con vistas al futuro, así como el conjunto de publicaciones relevantes derivadas de este trabajo.

6.1. CONCLUSIONES GENERALES

La aportación más importante de esta Tesis la constituye el cumplimiento de su objetivo principal, pues ha sido posible con la aplicación no supervisada y basada en conocimiento del Análisis Semántico Multidimensional en la Resolución de Ambigüedad Semántica de las Palabras, superar los resultados de sistemas actuales de similar condición y además lograr aplicar este análisis en el área de la Minería de Opiniones con resultados relevantes.

Esto ha sido posible, tras haber detectado que los sistemas de WSD han carecido de algún modo de multidimensionalidad semántica y que por lo general el *baseline* MFS siempre ha obtenido relevantes resultados. Estos aspectos hacen necesaria la creación de métodos de Resolución de Ambigüedad Semántica de las Palabras, capaces de comprobar si sería posible mejorar el estado actual de sistemas no supervisados y basados en conocimiento. En esta Tesis se han aplicado métodos sin supervisión, porque el principal problema de los sistemas supervisados, reside en la escasez de corpus anotados manualmente para su entrenamiento. Aunque es evidente, según los resultados obtenidos en las competiciones científicas analizadas, que este tipo de sistemas (incluye además a los débilmente supervisados) obtienen mejores resultados que los no supervisados (incluye los basados en conocimiento).

Para lograr el desarrollo de métodos que incorporen semántica multidimensional en sus procesos de WSD, se ha creado y evaluado satisfactoriamente un recurso semántico multidimensional (ISR-WN), capaz de ofrecer a sistemas de PLN diferentes ópticas de análisis textual. La concepción de esta herramienta ha sido posible, luego de haber estudiado los distintos recursos léxicos utilizados para el desarrollo de los métodos de WSD, donde el principal problema en el uso de estos recursos es su descentralización. A pesar de que la mayoría se basa en las relaciones internas de WN, no todos comparten una interfaz común que pueda proporcionar información de forma cohesionada. Aunque algunos autores han desarrollado redes semánticas integradoras, en su mayoría favorecen más la unificación lingüística que la conceptual.

Tras la creación del ISR-WN, se han desarrollado varios métodos de WSD basados en el Análisis Semántico Multidimensional. Inicialmente, se ha realizado un análisis desde múltiples dimensiones semánticas, en concreto construyendo Árboles Semánticos Relevantes (RST) de los textos. Esta primera variante, ha demostrado la viabilidad de su concepción, al obtener medianos aciertos con la interrelación de múltiples recursos. Como una mejora de RST se ha desarrollado RST+Frec. Esta utiliza los múltiples conceptos semánticos obtenidos por RST y aplica una medida que es capaz de evaluar los valores de frecuencias de sentidos conjuntamente con los de relevancia de RST. La nueva variante, ha logrado superar ampliamente a la original, al *baseline* MFS asociado a la frecuencia utilizada y además es capaz de colocarse a la cabeza de los resultados alcanzados por sistemas de WSD sin supervisión.

Otras de las propuestas desarrolladas se basan en grafos de conocimiento con la combinación del modelo *N-Cliques* en WSD. Esta obtiene sub-grafos que agrupan los conceptos más fuertemente enlazados. Aunque se han desarrollan dos variantes de *N-Cliques* (*N-Cliques*+RV y *N-Cliques* + RST) enfocadas en la creación del grafo de desambiguación, ambas aproximaciones han obtenido similares resultados, pudiendo ubicarse entre los puestos intermedios del *ranking* de Senseval-2.

Conocidas las deficiencias de *N-Cliques,* se ha introducido el método Ppr+Frec (basado en *Personalizing PageRank*, propuesta carente esencialmente de multidimensionalidad y frecuencia de sentidos). Este ha logrado obtener los mejores resultados entre los reportados para sistemas sin supervisión incluso por encima de RST+Frec, pudiendo resolver la ambigüedad semántica de los adverbios en un 100% al aplicarlo sobre el corpus de Senseval-3.

Después de haber experimentado notables mejoras en la resolución del Análisis Semántico Multidimensional sobre texto, se ha querido aplicar alguno de estos tipos especiales de resolución de ambigüedad en otras áreas del PLN. Por ejemplo en Minería de Opiniones. Para ello se ha planteado clasificar los textos según polaridades, presencia de opinión y/o pertenencia a un tópico o frase que lo describa.

Inicialmente, se ha aplicado la aproximación Senti-RST. Esta se basa en la construcción de Árboles Semánticos Relevantes capaces de describir a nivel conceptual los textos, en Árboles Conceptuales Positivos y Negativos. Senti-RST ha obtenido unos resultados, que ubicarían la propuesta entre los cuatro primeros puestos del *ranking* de la competición MOAT. Este aspecto es de suma relevancia, debido a que el objetivo de Senti-RST ha sido analizar la influencia del Análisis Semántico Multidimensional en el área de la Minería de Opiniones, obteniendo resultados prometedores en relación con una primera aproximación.

Por otra parte, se han propuesto varias aproximaciones de un método de Clasificación Textual, que se enfrentan a una de las problemáticas de esta tarea (la corta longitud del texto a analizar). La mejor de las aproximaciones de TC aplicada en opiniones, ha obtenido una precisión cercana al 60%. Aunque estos resultados difieren mucho de los obtenidos por los métodos tradicionales de TC, resulta de mucha ayuda poseer una propuesta que sea capaz de ser aplicada en textos cortos sin necesidad de aprendizaje.

La conclusión general de esta Tesis es que en la categoría de sistemas no supervisados al aplicar métodos basados en el conocimiento ofrecido por el Análisis Semántico Multidimensional, se obtienen los mejores resultados entre los sistemas de esta condición. Además otra conclusión muy relevante es que este tipo de análisis se puede aplicar al área de la Minería de Opiniones con resultados prometedores.

6.2. TRABAJOS FUTUROS

Aunque se han creado variadas producciones científicas a lo largo de todo el trabajo presente en la Tesis, estas han sentado las bases para la generación de nuevas propuestas futuras. Respecto al recurso ISR-WN, se propone su enriquecimiento incorporándole XWN1.7 y XWN3.0, con el fin de aplicar aproximaciones como Ppr+Frec sin la necesidad de recopilar las informaciones que ofrecen sus interconexiones de modo no centralizado. Teniendo en cuenta los resultados relevantes obtenidos por Senti-RST al combinar conceptualizaciones con polaridades, se propone además la incorporación de recursos como Micro-WNOp[77](corpus etiquetado a nivel de sentimientos compuesto por 1105 *synsets* de WN). Quedaría también para futuras proyecciones el alineamiento de ISR-WN con WN's en diferentes idiomas distintos del inglés, con el objetivo de emplear todas las investigaciones que en esta Tesis se defienden con carácter multilingüe.

Con el objetivo de lograr que ISR-WN incorpore cada vez más conocimiento, se propone como trabajo futuro, integrar automáticamente cualquier ontología, estudiando un modelo capaz

[77] http://www.unipv.it/micrownop

de fusionarlas al recurso, mediante el alineamiento automático entre *synsets* de WN y ontologías en formato RDF. Este proceso debe tener en consideración la información contextual que circunda a cada etiqueta de la ontología y la de los respectivos *synsets* de la palabra que le corresponde.

Nuevas aproximaciones de WSD quedan abiertas ahora que se ha comprobado que la interacción de frecuencia de sentidos con la multidimensionalidad de ISR-WN, se obtienen relevantes resultados. Por ejemplo, quedaría pendiente incorporar al proceso de *N-Cliques,* informaciones de importancia de sentidos que puedan ser obtenidas o bien por *PageRank* o por el simple uso de frecuencias de sentidos. También sería muy interesante realizar un análisis en textos anotados, para observar el comportamiento semántico existente entre cada sentido que se hace acompañar por otros de diferentes palabras en un corpus. Este comportamiento semántico basado en los caminos existentes entre sentidos de las palabras mediante el uso de ISR-WN (similar al ejemplo de la Figura 30), revelaría características claves que se cumplen a la hora de concebirse frases bien elaboradas.

Aún más propuestas del Análisis Semántico Multidimensional se pudieran aplicar en otras esferas del PLN. Por ejemplo en Clasificación de Textos sería muy interesante representar todas las palabras de los textos, agrupados con *N-Cliques.* De esta forma se pudieran concentrar en diferentes conjuntos la mayor cantidad de sentidos provenientes de textos similares y de la misma categoría.

6.3. PRODUCCIÓN CIENTÍFICA

A continuación se enumeran las publicaciones que han sido el resultado del trabajo de investigación. Estas se agrupan por año y además se incluye una referencia respecto a la participación en una competición internacional.

2011

1. **Gutiérrez Yoan**, Fernández Antonio, Montoyo Andrés and Vázquez Sonia 2011a Enriching the Integration of Semantic Resources based on WordNet. *Revista del Procesamiento del Lenguaje Natural*. **47** pp 249-57.
2. **Gutiérrez Yoan**, Vázquez Sonia and Montoyo Andrés 2011b Improving WSD using ISR-WN with Relevant Semantic Trees and SemCor Senses Frequency. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, (Hissar, Bulgaria: RANLP 2011) pp 233-9.
3. **Gutiérrez Yoan**, Vázquez Sonia and Montoyo Andrés 2011c Sentiment Classification Using Semantic Features Extracted from WordNet-based Resources. In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (*WASSA 2.011*), (Portland, Oregon: *Association for Computational Linguistics* (*ACL*)).
4. **Gutiérrez Yoan**, Vázquez Sonia and Montoyo Andrés 2011d Word Sense Disambiguation: A Graph-Based Approach Using *N-Cliques* Partitioning Technique. *In*: *16th International Conference on Applications of Natural Language to Information Systems* (*NLDB 2011*). pp 112-24.
5. Fernández Antonio, Díaz Josval, **Gutiérrez Yoan** and Muñoz Rafael 2011a An Unsupervised Method to Improve Spanish Stemmer. *In: 16th International Conference on Applications of Natural Language to Information Systems* (*NLDB 2011*). pp 221-4.
6. Fernández Antonio, **Gutiérrez Yoan**, Pérez Abel, García Yaniseth and Miranda Lisandra 2011b Extracción de Información en Documentos no Etiquetados del Entorno Educacional. *In: XII Simposio Internacional De Comunicación Social. Centro de Lingüística Aplicada Santiago de Cuba*, (Cuba), pp 970-4.
7. Pérez Abel, Fernández Antonio, **Gutiérrez Yoan** and Alfonso Yeiniel 2011 Generación de Expresiones Regulares para La Creación de Reglas en Aplicaciones de PLN. *In: XII Simposio Internacional De Comunicación Social. Centro de Lingüística Aplicada Santiago de Cuba*, (Cuba), pp 881-5.

2010

8. **Gutiérrez Yoan**, Fernández Antonio, Montoyo Andrés and Vázquez Sonia 2010a Integration of semantic resources based on WordNet. *Revista del Procesamiento del Lenguaje Natural.* **45** pp 161-88.
9. **Gutiérrez Yoan**, Fernández Antonio, Montoyo Andrés and Vázquez Sonia 2010b UMCC-DLSI: Integrative resource for disambiguation task. In: *Proceedings of the 5th International Workshop on Semantic Evaluation,* (Uppsala, Sweden: *Association for Computational Linguistics* (*ACL*)) pp 427-32.

PARTICIPACIÓN EN COMPETICIONES CIENTÍFICAS

10. Participación en la competición internacional Semeval 2010, Equipo: UMCC-DLSI. Tarea 17 "*All-words Word Sense Disambiguation on a Specific Domain* (*WSD-domain*)". Patrocinador: *Association for Computational Linguistic.* Evento*: Evaluation Exercises on Semantic Evaluation - ACL SigLex event.*

ACRÓNIMOS

- PLN: Procesamiento del Lenguaje Natural
- AI: *Artificial Intelligence*
- TPD: Tecnologías de Procesamiento de Datos
- TPLN: Tecnologías de Procesamiento del Lenguaje Natural
- WSD*: Word Sense Disambiguation*
- AC: *Affective Computing*
- POS*: Part of Speech*
- UML: *Unified Modeling Language*
- MRDs: *Machine-readable dictionaries*
- LDOCE: *Longman Dictionary of Contemporary English*
- BNC: *British National Corpus*
- DSO*: Defence Science Organisation*
- KIF: *Knowledge Interchange Format*
- XOL: *Ontology Exchange Language*
- OML: *Ontology Markup Language*
- RDF / RDFS : *Resource Description Framework Schema Language*
- OIL: *Ontology Interchange Language*
- OWL :*Ontology Web Language*
- URI : *Uniform Resource Identifier*
- XML : *Extensible Markup Language*
- WN: WordNet
- MRC*: Multilingual Central Repository*
- WND: WordNet *Domains*
- WNA: WordNet *Affects*
- SFC: *Subject Field Codes*
- SUMO :*Suggested Upper Merged Ontology*
- SC : *Semantic Classes*
- BLC: *Base Level Concepts*
- SWN: SentiWordNet
- XWN: *eXtended* WordNet
- MWN : MultiWordNet
- EWN : EuroWordNet
- *K*NN :*k-nearest neighbor*
- SVM: *Support Vector Machines*
- SVD*: Singular Value Decomposition*
- LCS: *Lowest Common Subsumer*
- MI *: Mutual Information*
- AR: *Association Ratio*
- RV*: Reuters Vector*
- UTN: Universidad del Norte de Texas
- RLSC*: Regularized Least Square Classification*
- ACL : *Association for Computational Linguistics*
- MFS*: Most Frequency Senses*
- RLSC: *Regularized Least Square Classification*
- SemEval: *Semantic Evaluations*
- ISR-WN*: Integration of Semantic Resources based in WordNet*

- RANLP: *Recent Advances in Natural Language Processing*
- NLDB: *Natural Language to Data Bases*
- LKB: *Lexical knowledge Base*
- RST *: Relevant Semantic Trees*
- RST+Frec: *Relevant Semantic Trees combinados con Frecuencia de Sentidos*
- *N-Cliques*+RST*: N-Cliques combinado con Relevant Semantic Trees*
- *N-Cliques*+RV*: N-Cliques combinado con Reuters Vector*
- Pr*: PageRank*
- Ppr*: Personalizing PageRank*
- Ppr+Frec*: Personalizing PageRank combinado con Frecuencias de sentidos*
- BFS *:Breath First Search*
- QA *: Question Answering*
- OQA: *Opinion Question Answering*
- MOAT*: Multilingual Opinion Analysis Task*
- TC: *Text Categorization*
- Senti-RST*: Relevant Semantic Trees Sentimental*
- PST: *Positive Semantic Tree*
- NST: *Negative Semantic Tree*
- PosA*: Positive Association*
- NegA*: Negative Association*
- MSF-TC*: Multidimensional Semantic Features to Text Classification*

REFERENCIAS BIBLIOGRÁFICAS

Achen, C. H. (2008) *Interpreting and Using Regression,* Princeton University.

Agirre , E., & German Rigau (1996) Word Sense Disambiguation using Conceptual Density.

Agirre, E. & De_Lacalle, O. L. (2007) UBC-ALM: Combining k-NN with SVD for WSD. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Agirre, E. & Edmonds, P. (2006) Algorithms and Applications. *Text, Word Sense Disambiguation.*

Agirre, E., Lacalle, O. L. D., Fellbaum, C., Hsieh, S.-K., Tesconi, M., Monachini, M., Vossen, P. & Segers, R. (2010) SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Los Angeles, California, Association for Computational Linguistics.

Agirre, E., Márquez, L. & Wicentowski, R. (2007) Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). *Association for Computational Linguistics.* Prague, Czech Republic.

Agirre, E. & Martinez, D. (2001) Learning class-to-class selectional preferences. *Proceedings of the 5th Conference on Computational Natural Language Learning (CoNLL).* Toulouse, France.

Agirre, E. & Rigau, G. (1996) Word sense disambiguation using Conceptual Density. *Proceedings of the 16th conference on Computational linguistics - Volume 1*. Copenhagen, Denmark, Association for Computational Linguistics.

Agirre, E. & Soroa, A. (2009) Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009).* Athens, Greece.

Aho, Hopcroft & Ullman (1974) The Design and Analysis of Computer Algorithms. *Addison, Wesley*.

Alba, R. D. (1973) A Graph-Theoretic Definition of a Sociometric Clique. *Journal of Mathematical Sociology,* 3**,** 113-126.

Alvarez, M. & All, E. (1998) *Matemática numérica,* La Habana, Editorial Félix Varela.

Alvarez, M. & Et.Al. (2004) *Matemática Numérica,* La Habana, Editorial Felix Varela.

Andrews, P. & Rajman, M. (August 2004) Thematic Annotation: extracting concepts out of documents. Lausanne, School of Computer & Communication Sciences. Swiss Federal Institute of Technology.

Atkins, S. (1993) Tools for Computer-aided Corpus Lexicography: the Hector Project. *Acta Linguistica Hungarica,* 41**,** 5--72.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. & Padró, M. (2006) FreeLing 1.3: Syntactic and semantic services in an opensource NLP library. *Proceedings of LREC'06.* Genoa, Italy.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B. & Vossen, P. (2004) The MEANING Multilingual Central Repository. *Proceedings of the Second International Global WordNet Conference (GWC'04).* . Brno, Czech Republic.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010) SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. IN 2010, L. (Ed.) *7th Language Resources and Evaluation Conference.* Valletta, MALTA.

Balahur, A., Boldrini, E., Montoyo, A. & Martinez-Barco, P. (2010) The OpAL System at NTCIR 8 MOAT. *Proceedings of NTCIR-8 Workshop Meeting.* Tokyo, Japan.

Balahur, A. & Montoyo, A. (2009) A Semantic Relatedness Approach to Classifying Opinion from Web Reviews. *Procesamiento del Lenguaje Natural,* 42**,** 47-54.

Balasundaram, B., Butenko, S., Hicks, I. V. & Sachdeva, S. (2006) Clique relaxations in Social Network Analisis: The Maximun k-plex Problem.

Banerjee, S. & Pedersen, T. (2002) Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet.

Banfield, A. (1982) Unspeakable sentences: Narration and Representation in the Language of Fiction. *Routledge and Kegan Paul*.

Barzilay, R. & Elhadad, M. (1997) Using lexical chains for text summarization. *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization* Madrid, Spain.

Bentivogli, L., Forner, P., Magnini, B. & Pianta, E. (2004) Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing. *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources".* Geneva, Switzerland. .

Bernhard, E. B., Isabelle, M. G. & Vladimir, N. V. (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory.* Pittsburgh, Pennsylvania, United States, ACM.

Boguslavsky, I. M., Iomdin, L. L., Lazursky, A. V., Mityushin, L. G., Sizov, V. G., Kreydlin, L. G. & Berdichevsky, A. S. (2005) Interactive Resolution of Intrinsic and Translational Ambiguity in a Machine Translation System. *CICLing 2005.* A. Gelbukh (ed.), Springer-Verlag Berlin Heidelberg Lecture notes in computer science.

Brill, E. (1995) Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. MIT Press.

Brin, S. & Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 30**,** 107-117.

Buscaldi, D. & Rosso, P. (2007) UPV-WSD : Combining different WSD Methods by means of Fuzzy Borda Voting. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Cai, J. F., Lee, W. S. & Teh, Y. W. (2007) NUS-ML: Improving Word Sense Disambiguation Using Topic Features. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Cavique, L., Mendes, A. B. & Santos, J. M. A. (2009) An Algorithm to Discover the k-Clique Cover in Networks. *Progress in Artificial Intelligence.* . Springer Berlin / Heidelberg.

Cilibrasi, R. L. & Vitányi, P. M. B. (2007) The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering,* VOL. 19, NO 3.

Clark, B. N., Colbourn, C. J. & Johnson, D. S. (1990) *Unit Disk Graphs,* North-Holland, Elsevier Science Publishers B.V.

Clough, P. (2005) Caption and Query Translation for Cross-Language Image Retrieval. *C. Peters et al. (Eds.): CLEF 2004***,** pp. 614–625.

Corcho, O., Fernández-López, M., Gómez-Pérez, A., Angele, J., Bechhofer, S., Domingue, J., Légér, A., Missikoff, M., Motta, E., Musen, M., Noy, N. F., Sure, Y., Taglino, F., Mcguiness, D., Ramos, J. A., Stumme, G., Bouillon, Y., Stutt, A., Handschuh, S.,

López, A., Maier-Collin, M., Christophides, V., Plexousakis, D., Magkanaraki, A., Ahn, T. T. & Karvounarakis, G. (2002) Deliverable 1.3: A survey on ontology tools. *OntoWeb: Ontology-based information exchange for knowledge management and electronic commerce.* UPM.

Cormen, T. H., Leiserson, C. E. & Rivest, R. L. (1990) Introduction to Algorithms. MA/McGraw-Hill, New York, MIT press, Cambridge.

Cotton, S., Edmonds, P., Kilgarriff, A. & Palmer, M. (2001) English All word. IN LINGUISTICS, A. F. C. (Ed.) *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems.* Toulouse, France, Association for Computational Linguistics.

Cowie, J., Guthrie, J. & Guthrie, L. (1992) Lexical Disambiguation using Simulated Annealing. In Proceedings of the International Conference on Computational Linguistics.

Cuadros, M. & Rigau, G. (2006) Quality assessment of large scale knowledge resources. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Sydney, Australia., Association for Computational Linguistics

Chan, Y. S., Ng, H. T. & Zhong, Z. (2007) NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Chen, H., Chung, Y.-M., Ramsey, M. & Yang, C. C. (1998) A Smart Itsy Bitsy Spider for the Web. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE.,* 49**,** 604–618.

Chen, P., Ding, W., Choly, M. & Bowes, C. (2010) Word Sense Disambiguation with Automatically Acquired Knowledge. *IEEE Intelligent Systems*.

Chi, Z. & Yan, H. (1995) Feature Evaluation and Selection Based on an Entropy Measure with Data Clustering. *Optical Engineering,* 34.

Chklovski, T. & Mihalcea, R. (2002) Building a sense tagged corpus with open mind word expert. *Proceedings of the ACL-02 workshop on Word sense disambiguation.* NJ, USA., Association for Computational Linguistics, Morristown.

Church, K. W. & Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* Volume 16, Number 1.

David, Y. (1994) Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Las Cruces, New Mexico, Association for Computational Linguistics.

Decadt, B., Hoste, V., Daelemans, W. & Bosch, A. V. D. (2004) GAMBL, genetic algorithm optimization of memory-based WSD. IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Díaz, A. Algunos Algoritmos Sobre Gráficas. *Análisis y Diseño de Algoritmos.* Departamento de Ingeniería Eléctrica CINVESTAV-IPN.

Diederich, J., Kindermann, J. R., Leopold, E. & Paass, G. (2003) Authorship Attribution with Support Vector Machines. *Applied Intelligence,* 19**,** 109-123.

Dijkstra, E. W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik.*

Dorr, B. J. & Castellón, M. A. M. A. I. (1997) Spanish EuroWordNet and LCS-Based Interlingual MT. *AMTA/SIG-IL First Workshop on Interlinguas.* San Diego, CA.

Ellman, J., Klincke, I. & Tait, J. (2000) Word Sense Disambiguation by Information Filtering and Extraction. *Computers and the Humanities* 34**,** 127–134.

Ercan, G. & Cicekli, I. (2007) Using lexical chains for keyword extraction. *Information Processing & Management,* 43**,** 1705-1714.

Escudero, G., Marquez, L. & Rigau, G. (2000a) Boosting Applied to Word Sense Disambiguation. *PROCEEDINGS OF THE 12TH EUROPEAN CONFERENCE ON MACHINE LEARNING.*

Escudero, G., Márquez, L. & Rigau, G. (2000b) Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited. *ECAI.*

Esuli, A. & Sebastiani, F. (2006) SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. IN 2006, L. (Ed.) *Fifth international conference on Languaje Resources and Evaluation* Genoa - ITaly.

Fellbaum, C. (1998) *WordNet. An Electronic Lexical Database,* University of Cambridge.

Fernández-Amorós, D., Gonzalo, J. & Verdejo, F. (2009) The UNED systems at Senseval-2. *CoRR.*

Fernández, P. (1996) Determinación del tamaño muestral.

Floyd, R. W. (1963) Algothm 97: Shortest path. *C. ACM,* 5**,** 345.

Forner, P. (2005) WordNet Domains 2.0. *ITC-irst, Povo-Trento, Italy*.

Freund, Y. & Schapire, R. E. (1999) A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence,* 14**,** 771-780.

Friedkin, N. E. (1984) Structural Cohesion and Equivalence Explanations of Social Homogeneity. *Sociological Methods & Research,* 12**,** 235-261.

Gale, W. A., Church, K. W. & Yarowsky, D. (1992a) Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings of the 30th annual meeting on Association for Computational Linguistics.* Newark, Delaware, Association for Computational Linguistics.

Gale, W. A., Church, K. W. & Yarowsky, D. (1992b) A Method for Disambiguating Word Senses in a Large Corpus. *Common Methodologies in Humanities Computing and Computational* 26**,** 415-439

Galley, M. & Mckeown, K. (2003) Improving word sense disambiguation in lexical chaining. *Proceedings of the 18th international joint conference on Artificial intelligence.* Acapulco, Mexico., Morgan Kaufmann Publishers Inc.

Genesereth, M. R. & Fikes, R. E. (1992) Knowledge Interchange Format. IN UNIVERSITY, C. S. D. S. (Ed.) version 3.0 Reference Manual ed. Stanford, Computer Science Department.

Gerard, S. (1989) *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Addison-Wesley Longman Publishing Co., Inc.

Gildea, D. & Jurafsky, D. (2002) Automatic labeling of semantic roles. MIT Press.

Gliozzo, A., Strapparava, C. & Dagan, I. (2004) Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language.*

Gómez-Allende, D. M. (1993) *Reconocimiento de Formas y Vision Artificial,* Madrid : RA-MA.

Grozea, C. (2004) Finding optimal parameter settings for high performance word sense disambiguation. IN EDMONDS, R. M. A. P. (Ed.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Gruber, T. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5(2):199-220*.

Gutiérrez, Y. (2010) Resolución de ambigüedad semántica mediante el uso de vectores de conceptos relevantes. *Departamento de Informática*. Matanzas, UMCC.

Gutiérrez, Y., Fernández, A., Montoyo, A. & Vázquez, S. (2010a) Integration of semantic resources based on WordNet. IN 2010, S. (Ed.) *XXVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.* Universidad Politécnica de Valencia, Valencia, SEPLN 2010.

Gutiérrez, Y., Fernández, A., Montoyo, A. & Vázquez, S. (2010b) UMCC-DLSI: Integrative resource for disambiguation task. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden, Association for Computational Linguistics.

Gutiérrez, Y., Fernández, A., Montoyo, A. & Vázquez, S. (2011a) Enriching the Integration of Semantic Resources based on WordNet. *Procesamiento del Lenguaje Natural,* 47**,** 249-257.

Gutiérrez, Y., Vázquez, S. & Montoyo, A. (2011b) Improving WSD using ISR-WN with Relevant Semantic Trees and SemCor Senses Frequency. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011.* Hissar, Bulgaria, RANLP 2011 Organising Committee.

Gutiérrez, Y., Vázquez, S. & Montoyo, A. (2011c) Sentiment Classification Using Semantic Features Extracted from WordNet-based Resources. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011).* Portland, Oregon., Association for Computational Linguistics.

Gutiérrez, Y., Vázquez, S. & Montoyo, A. (2011d) Word Sense Disambiguation: A Graph-Based Approach Using N-Cliques Partitioning Technique. *NLDB'11.* Alicante. Spain.

Halliday, M. A. K. & Hasan, R. (1976) *Cohesion in English,* London, U.K., Longman Group Ltd.

Hatzivassiloglou, Vasileios & Wiebe, J. (2000) Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *International Conference on Computational Linguistics (COLING-2000).*

Haveliwala, T. H. (2002) Topic-sensitive pagerank. *WWW '02: Proceedings of the 11th international conference on World Wide Web.* New York, NY, USA., ACM.

Haveliwala, T. H. (2003) Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering,* 15**,** 784-796.

Hawkins, P. & Nettleton, D. (2000) Large Scale WSD Using Learning Applied to SENSEVAL. *Computers and the Humanities,* 34 135–140.

Hedlund, T., Keskustalo, H., Pirkola, A., Sepponen, M. & Järvelin, K. (2001) Bilingual Tests with Swedish, Finnish, and German Queries: Dealing with Morphology, Compound Words, and Query Structure. *C. Peters (Ed.): CLEF 2000, LNCS 2069***,** pp. 210-223.

Hinrich, S. (1998) Automatic word sense discrimination. MIT Press.

Hinrichschütze (1997) Ambiguity Resolution in Language Learning. Stanford, California, CSLI.

Hirst, G. & St-Onge, D. (1998) Lexical chains as representations of context for the detection and correction of malapropisms. IN FELLBAUM, C. (Ed.) *WordNet: An Electronic Lexical Database.* Cambridge, MA., The MIT Press.

Hovy, E. (2003) Using an Ontology to Simplify an Data Acces. *Communications of the ACM,* Vol. 46, No. 1.

Ide, N. & Véronis, J. (1998) Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics. MIT Press,* 24**,** 2--40.

Ion, R. & Stefanescu, D. (2010) RACAI: Unsupervised WSD Experiments @ SemEval-2, Task 17. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden, Association for Computational Linguistics.

Ion, R. & Tufis, D. (2007) RACAI: Meaning Affinity Models. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Izquierdo, R. (2010) Una Aproximación a la Desambiguación del Sentido de las Palabras Basada en Clases Semánticas y Aprendizaje Automático. *Departamento de Lenguajes y Sistemas Informáticos.* Alicante, Universidad de Alicante.

Izquierdo, R., Suárez, A. & Rigau, G. (2007) A Proposal of Automatic Selection of Coarse-grained Semantic Classes for WSD. *Procesamiento del Lenguaje Natural,* 39**,** 189-196.

Izquierdo, R., Suárez, A. & Rigau, G. (2010) GPLSI-IXA: Using Semantic Classes to Acquire Monosemous Training Examples from Domain Texts *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden, Association for Computational Linguistics.

Jiang, J. J. & Conrath, D. W. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics (ROCLING X).* Taiwan.

Khapra, M. M., Shah, S., Kedia, P. & Bhattacharyya, P. (2010) Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. *5th International Conference on Global Wordnet (GWC2010).* Indian Institute of Technology, Bombay, Global Wordnet Association

Kilgariff, A. & Palmer, M. (1998) Proceedings of the Pilot SensEval. *Senseval-1*. Hermonceux Castle, Sussex, UK., Association for Computational Linguistics.

Kilgarriff, A. & Rosenzweig, J. (2000) Framework and Results for English SENSEVAL *Computers and the Humanities.* Springer.

Kim, S.-M. & Hovy, E. (2005) Identifying Opinion Holders for Question Answering in Opinion Texts. *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains.*

Kim, S.-M. & Hovy, E. (2006) Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. *In Proceedings of workshop on sentiment and subjectivity in text at proceedings of the 21st international conference on computational linguistics/the 44th annual meeting of the association for computational linguistics (COLING/ACL 2006).* Sydney, Australia.

Klema, V. & Laub, A. (1980) The singular value decomposition: Its computation and some applications. *Automatic Control, IEEE Transactions on,* 25**,** 164-176.

Klyne, G. & J.Carroll, J. (2006) Resource Description Framework (RDF): Concepts and Abstract Syntax. IN MCBRIDE, B. (Ed.).

Kose, F., Weckwerth, W., Linke, T. & Feihn, O. (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics.*

Krovetz, R. (1998) More than One Sense Per Discourse. *Proceedings of the ACL-SIGLEX Workshop.*

Kulkarni, A., Khapra, M., Sohoney, S. & Bhattacharyya, P. (2010) CFILT: Resource Conscious Approaches for All-Words Domain Specific WSD *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden, Association for Computational Linguistics.

Kwok, K. L. (1999) English-Chinese Cross-Language Retrieval based on a Translation Package. *MT Summit VII Workshop: Machine Translation for Cross Language IR.* City University of New York. Flushing, NY 11367, USA, Computer Science Dept., Queens College.

Lai, G.-H., Huang, J.-W., Gao, C.-P. & Tsai, R. T.-H. (2010) Enhance Japanese Opinionated Sentence Identification using Linguistic Features: Experiences of the IISR Group at NTCIR-8 MOAT Task. *Proceedings of NTCIR-8 Workshop Meeting.* Tokyo, Japan.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998) An Introduction to Latent Semantic Analysis. *Discourse Processes.*

Laparra, E., Rigau, G. & Cuadros, M. (2010) Exploring the integration of WordNet and FrameNet. *Proceedings of the 5th Global WordNet Conference (GWC'10).* Mumbai, India.

Leacock, C. & Chodorow, M. (1998) Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics.*

Leacock, C., Towell, G. & Voorhees, E. (1993) Corpus-based statistical sense resolution. *Proceedings of the workshop on Human Language Technology.* Princeton, New Jersey, Association for Computational Linguistics.

Lesk, M. E. (1986) Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. *Proceedings of the ACM SIGDOC Conference.* Toronto, Ontario.

Li, J. & Sun, M. (2007) Scalable Term Selection for Text Categorization. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Prague, Association for Computational Linguistics.

Lin, D. (1998a) Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational linguistics (COLING).* Montreal, Canada.

Lin, D. (1998b) An Information-Theoretic Definition of Similarity. *Proceedings of International Conference on Machine Learning.* Madison, Wisconsin.

Luce, R. D. (1950) Connectivity and generalized cliques in sociometric group structure. *Psychometrika.*

Luce, R. D. & Perry, A. D. (1949) A Method of Matrix Analysis of Group Structure. *Psychometrie.*

Luisa Bentivogli, P. F., Bernardo Magnini, Emanuele Pianta (2005) Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing. *ITC-irst – Istituto per la Ricerca Scientifica e Tecnologica Via Sommarive 18, Povo – Trento, Italy, 38050*.

Mackay, D. (2003) Chapter 20. An Example Inference Task: Clustering. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press.

Magnini, B. & Cavaglia, G. (2000) Integrating Subject Field Codes into WordNet. *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-2000).*

Magnini, B., Satrapparava, C., Pezzulo, G. & Gliozzo, A. (July 2002) The Role of Domains Informations in Word Sense Disambiguatios. Treto, Cambridge University Press.

Magnini, B., Strapparava, C., Pezzulo, G. & Gliozzo, A. (2002) Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet. *Proceedings of the First International WordNet Conference.* Mysore, India.

Magnini, B., Strapparava, C., Pezzulo, G. & Gliozzo, A. (2008) Using Domain Information for Word Sense Disambiguation. *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology (icetet 2008).* Nagpur, India.

Martín, T., Balahur, A., Montoyo, A. & Pons, A. (2010) Word sense disambiguation in opinion mining: Pros and cons. *Proc. of CICLing'10.* Madrid,Spain.

Martinez, D. & Agirre, E. (2000) One sense per collocation and genre/topic variations. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13.* Hong Kong, Association for Computational Linguistics.

Martinez, D., Baldwin, T., Agirre, E. & De_Lacalle, O. L. (2007) UBC-UMB: Combining unsupervised and supervised systems for all-words WSD. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Mc.Carthy, D., Koeling, R., Weeds, J. & Carroll, J. (2004) Finding Predominant Word Senses in Untagged Text. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04).* Barcelona, Spain.

Mccarthy, J. (1959) Programs with Common Sense. *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory.* London, U.K., Her Majesty's Stationery Office. Reprinted in John McCarthy. Formalizing Common Sense: Papers by John McCarthy. Ablex Publishing Corporation, 1990.

Mcculloch, W. S. & Pitts, W. (1988) A logical calculus of the ideas immanent in nervous activity. *Neurocomputing: foundations of research.* MIT Press.

Mezquita, Y. L. (2008) Recuperación de Información con Resolución de Ambigüedad del Sentido de las Palabras para el Español. *Red de Revistas Científicas de América Latina y el Caribe, España y Portugal,* 11**,** 288-300.

Mihalcea, R. (2005) Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *Proceedings of HLT05.* Morristown, NJ, USA.

Mihalcea, R., Chklovski, T. & Kilgarrif, A. (2004) The Senseval-3 English lexical sample task. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Mihalcea, R. & Faruque, E. (2004) SenseLearner: Minimally supervised Word Sense Disambiguation for all words in open text. IN EDMONDS, R. M. A. P. (Ed.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Mihalcea, R. & Moldovan, D. I. (1998) Word Sense Disambiguation based on Semantic Density. *Content Visualization and Intermedia Representations (CVIR'98).*

Mihalcea, R. & Moldovan, D. I. (1999) A Method for Word Sense Disambiguation of Unrestricted Text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic.*

Mihalcea, R. & Moldovan, D. I. (2001) Pattern Learning and Active Feature Selection for Word Sense Disambiguation. *SensEval-2.*

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography, 3(4):235-244.*

Mokken, R. J. (1979) *Cliques, Clubs and Clans*, Amsterdam, Elsevier Scientific Publishing Company.

Moldovan, D. I. & Rus, V. (2001) Explaining Answers with Extended WordNet. *ACL*.

Molina, A., Pla, F., Segarra, E. & Moreno, L. (2002) Word Sense Disambiguation using Statistical Models and WordNet. IN LREC 2002 (Ed.) *Proceedings of 3rd International Conference on Language Resources and Evaluation.* Las Palmas de Gran Canaria, LREC2002.

Mongay, C. (2005) *Quimiometría*, Universidad de Valencia.

Montoyo, A. (2002) Desambiguación léxica mediante Marcas de Especificidad. *Dpto. Sistemas y Lenguajes Informáticos.* Alicante, Spain., Universidad de Alicante.

Montoyo Guijarro, A. (2002) Desambiguación léxica mediante Marcas de Especificidad. *Dpto. Sistemas y Lenguajes Informáticos.* Alicante, Spain., Universidad de Alicante.

Montoyo Guijarro, A. (2008) Uso y diseño de ontologías en procesamiento del lenguaje natural y la web semántica.

Mooney, R. J. (1996) Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning *Conference on Empirical Methods in Natural Language Processing (EMNLP 96).*

Naskar, S. K. & Bandyopadhyay, S. (2007) JU-SKNSB: Extended WordNet Based WSD on the English All-Words Task at SemEval-1. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Navigli, R. (2009) Word sense disambiguation: A survey. *ACM Comput. Surv.,* 41**,** 10:1--10:69.

Navigli, R. & Lapata, M. (2007) Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. *IJCAI.*

Navigli, R., Litkowski, K. C. & Hargraves, O. (2007) SemEval-2007 Task 07: Coarse-Grained English All-Words Task. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Navigli, R. & Velardi, P. (2004) Structural semantic interconnection: a knowledge-based approach to Word Sense Disambiguation IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Navigli, R. & Velardi, P. (2005) Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27**,** 1075-1086.

Niles, I. (2001) Mapping WordNet to the SUMO Ontology. *Teknowledge Corporation*.

Niles, I. & Pease, A. (2001) Origins of the IEEE Standard Upper Ontology. *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology.* Seattle, Washington, USA.

Niles, I. & Pease, A. (2003) Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology.

Oliveira, C. S. (2006) ALEXIA - Acquisition of Lexical Chains for Text Summarization. *Department of Computer Science.* University of Beira Interior.

Padilla, A. P. (2002a) Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe. Ambigüedad, variación terminológica y multilingüismo. *Sociedad Española para el Procesamiento de Lenguaje Natural*.

Padilla, A. P. (2002b) Un sistema interactivo y multilingüe de búsqueda textual basado en técnicas lingüísticas. *Departamento de Lenguajes y Sistemas Informáticos.* Universidad Nacional de Educación a Distancia.

Pang, B. & Lee, L. (2003) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting of the ACL.*

Pang, B., Lee, L. & Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceeding of the ACM Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Pease, A. (2007) Standard Upper Ontology Knowledge Interchange Format.

Pekar, V. & Krkoska, M. (2003) Weighting Distributional Features for Automatic Semantic Classification of Words. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03).*

Pianta, E., Bentivogli, L. & Girardi, C. (2002) MultiWordNet. Developing an aligned multilingual database. *Proceedings of the 1st International WordNet Conference*. Mysore, India.

Picard, R. (1995) Affective computing. *Technical report.* MIT Media Laboratory.

Popescu, M. (2004) Regularized Least-Squares classification for Word Sense Disambiguation. IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Pradhan, S. S., Loper, E., Dligach, D. & Palmer, M. S. (2007) SemEval-2007 Task-17: English Lexical Sample SRL and All Words. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Association for Computational Linguistics.

Preiss, J. (2006) A detailed comparison of WSD systems:an analysis of the system answers for the Senseval-2 English all words task. *Natural Language Engineering,* 12**,** 209–228.

Quinlan, J. R. (1986) Induction of Decision Trees. Kluwer Academic Publishers.

Ratnaparkhi, A. (1998) Maximum entropy models for natural language ambiguity resolution. University of Pennsylvania.

Reddy, S., Inumella, A., Mccarthy, D. & Stevenson, M. (2010) IIITH: Domain Specific Word Sense Disambiguation. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden. , Association for Computational Linguistics.

Resnik, P. (1993) Selection and Information: A Classbased Approach to Lexical Relationships. University of Pennsylvania.

Resnik, P. (1995) Using information content to evaluate semantic similarity. *Proceedings of the 14th International Joint Conference on Artificial Intelligence.* Montreal.

Richardson, S. D. (1997) Determining Similarity and Inferring Relations in a Lexical Knowledge Base. The City University of New York.

Rigau Claramunt, G. (1998) Automatic Acquisition of Lexical Knowledge from MRDs. Barcelona, Universidad Politécnica de Cataluña.

Ronald, L. R. (1987) Learning Decision Lists. Kluwer Academic Publishers.

Russell, S. & Norvigy, P. (1994) A Modern, Agent-Oriented Approach to Introductory Artificial Intelligence.

Salton, G. & Mcgill, M. J. (1986) *Introduction to Modern Information Retrieval,* New York, NY, USA, McGraw-Hill, Inc.

Salzberg, S. L. (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning,* 16**,** 235-240.

Samper, J. J. (2005) Ontologías para Servicios Web Semánticos de Información de Tráfico: descripción y herramientas de explotación. *Departamento de Informática.* Valencia, Spain., Universidad de Valencia.

Sanda M. Harabagiu, G. A. M., Dan I. Moldovan (1999) Wordnet 2-A morphologically and semantically enhanced resource. *SIGLEX99: Standardizing Lexical Resources.*

Santamaría, C. (2010) Collaboratively Authored Web Contents as Resources for Word Sense Disambiguation and Discovery. *Departamento de Lenguajes y Sistemas Informáticos.* Universidad Nacional de Educación a Distancia. Escuela Técnica Superior de IngenierÍia Informática.

Santiago, F. M. (2004) El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia. *Departamento de Lenguajes y Sistemas Informáticos.* Universidad Nacional de Educación a Distancia.

Sara, T. & Daniele, P. (2009) New features for FrameNet: WordNet mapping. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning.* Boulder, Colorado, Association for Computational Linguistics.

Schneider, G. (2005) Distributionalism: Document Classification, Word Sense Disambiguation, Word Similarity. *Computerlinguistik, und Lexikologie Institut für Informatik.*

Schütze, H. (1997) Ambiguity Resolution in Language Learning. Stanford, California, CSLI.

Sebastiani, F. (1999) A tutorial on automated text categorisation. *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence.* Buenos Aires, Argentina.

Ševčenko, M. (2003) Online Presentation of an Upper Ontology. CTU Prague, Dept of Computer Science.

Silber, H. G. & Mccoy, K. F. (2000) Efficient text summarization using lexical chains. *Proceedings of the 5th international conference on Intelligent user interfaces.* New Orleans, Louisiana, United States, ACM.

Sinha, R. & Mihalcea, R. (2007) Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007).* Irvine, CA.

Snyder, B. & Palmer, M. (2004) The English All Word Task. IN LINGUISTICS, A. F. C. (Ed.) *SENSEVAL-3: Third International Workshop on the evaluation of System of the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Soroa, A., Agirre, E., Lacalle, O. L. D., Bosma, W., Vossen, P., Monachini, M., Lo, J. & Hsieh, S.-K. (2010) Kyoto: An Integrated System for Specific Domain WSD. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden., Association for Computational Linguistics.

Strapparava, C., Gliozzo, A. & Giuliano, C. (2004) Pattern abstraction and term similarity for Word Sense Disambiguation: IRST at Senseval-3. IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Association for Computational Linguistics.

Strapparava, C. & Valitutti, A. (2004) WordNet-Affect: an affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).* Lisbon.

Tello, A. L. (2001) Ontologías en la Web Semántica. *Área de Lenguajes y Sistemas Informáticos.* Departamento de Informática. Escuela Politécnica. Avda de la Universidad s/n, 10071 Cáceres. Universidad de Extremadura. España.

Thorsten, J. (1998) Text Categorization with Support Vector Machines: Learning with many relevant features. *10th European Conference on Machine Learning.* Chemnitz, Germany, {ECML}-98.

Tran, A., Bowes, C., Brown, D., Chen, P., Choly, M. & Ding, W. (2010) TreeMatch: A Fully Unsupervised WSD System Using Dependency Knowledge on a Specific Domain. *Proceedings of the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden, Association for Computational Linguistics.

Tratz, S., Sanfilippo, A., Gregory, M., Chappell, A., Posse, C. & Whitney, P. (2007) PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague, Czech Republic, Association for Computational Linguistics.

Tsatsaronis, G., Vazirgiannis, M. & Androutsopoulos, I. (2007) Word sense disambiguation with spreading activation networks generated from thesauri. *IJCAI.*

Tseng, C.-J. & Siewiorek, D. P. (1986) Automated Synthesis of Data Paths in Digital Systems. *IEEE Trans. on CAD of Integrated Circuits and Systems,* 5**,** 379-395.

Valitutti, A., Strapparava, C. & Stock, O. (Eds.) (2004) *Developing Affective Lexical Resources,* ITC-irst, Trento, Italy, PsychNology Journal.

Vasilescu, F., Langlais, P. & Lapalme, G. (2004) Evaluating Variants of the Lesk Approach for Disambiguating Words. *Proceedings of LREC 2004.* Lisbonne.

Vázquez Pérez, S., Montoyo, A. & Rigau, G. (2004) Using Relevant Domains Resource for Word Sense Disambiguation. *in IC-AI'04. Proceedings of the International Conferenceon Artificial Intelligence.* Ed: CSREA Press. Las Vegas, E.E.U.U.

Vázquez, S. (2009) Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN. *Depto. de Lenguajes y Sistemas Informáticos.* Alicante, Spain., Universidad de Alicante.

Vázquez, S., Montoyo, A. & Rigau, G. (2004a) Using Relevant Domains Resource for Word Sense Disambiguation. *IC-AI'04. Proceedings of the International Conference on Artificial Intelligence.* Ed: CSREA Press. Las Vegas, E.E.U.U.

Vázquez, S., Romero, R., Suárez, A., Montoyo, A., Nica, I. & Martí, A. (2004b) The University of Alicante systems at Senseval-3. IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Veronis, J. (2004) HyperLex: lexical cartography for information retrieval. *Computer Speech & Language,* 18**,** 223--252.

Vilares, J. (2005) Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en español. *IV Edición de los Premios SEPLN a la Investigación en Procesamiento del Lenguaje Natural.*

Villarejo, L., Márquez, L., Agirre, E., Martínez, D., Magnini, B., Strapparava, C., Mccarthy, D., Montoyo, A. & Suárez, A. (2004) The "Meaning" system on the English all-words task. IN MIHALCEA, R. & EDMONDS, P. (Eds.) *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Barcelona, Spain, Association for Computational Linguistics.

Villarejo, L., Márquez, L. & Rigau, G. (2005) Exploring the construction of semantic class classiers for WSD. *Sociedad Española para el Procesamiento del Lenguaje Natural.*

Vossen, P. (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht, Kluwer Academic Publishers.

Vossen, P., Peters, W. & Gonzalo, J. (1999) Towards a Universal Index of Meaning. *proceedings of the ACL-99 Siglex workshop.* University of Maryland.

Warren, W. (1955) Translation. *Machine Translation of Languages*.

Widdows, D. & Dorow, B. (2002) A Graph Model for Unsupervised Lexical Acquisition. *The 19th International Conference on Computational Linguistics (COLING 2002).* Taipei, Taiwan.

Wiebe, J., Wilson, T. & Cardie, C. (2005) Annotating Expressions of Opinions and Emotions in Language. *Kluwer Academic Publishers.* Netherlands.

Wilks, Y. & Stevenson, M. (1996) The grammar of sense: Is word sense tagging much more than part-of-speech tagging? , University of Sheffield.

Wilks, Y. & Stevenson, M. (1998) Word sense disambiguation using optimised combinations of knowledge sources. *Coling-ACL*.

Wood, D. R. (1997) An algorithm for finding a maximum clique in a graph. *Operations Research Letters* 211-217.

Yarowsky, D. (1993) One sense per collocation. *Proceedings of the DARPA Workshop on Human Language Technology.* Princenton, NJ.

Yarowsky, D. (2000) Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities,* 34**,** 179-186.

Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C. & Wicentowski, R. (2001) The Johns Hopkins SENSEVAL2 System Descriptions. *Proceedings of SENSEVAL-2.*

Zouaq, A., Gagnon, M. & Ozell, B. (2009) A SUMO-based Semantic Analysis for Knowledge Extraction. *Proceedings of the 4th Language & Technology Conference.* Poznań, Poland.

Zubaryeva, O. & Savoy, J. (2010) Opinion Detection by Combining Machine Learning & Linguistic Tools *Proceedings of NTCIR-8 Workshop Meeting.* Tokyo, Japan.

Reunido el Tribunal que suscribe en el día de la fecha acordó otorgar, por a la Tesis Doctoral de Don/Dña. la calificación de .

Alicante de de

El Secretario,

El Presidente,

UNIVERSIDAD DE ALICANTE

CEDIP

La presente Tesis de D. _____ ha sido registrada con el nº _____ del registro de entrada correspondiente.

Alicante ___ de _____ de _____

El Encargado del Registro,