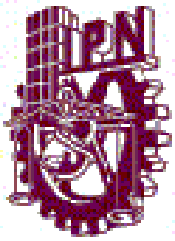
INSTITUTO POLITÉCNICO NACIONAL



CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Laboratorio de Lenguaje Natural
y Procesamiento de Texto

CONSTRUCCIÓN AUTOMÁTICA DE UN MODELO DE ESPACIO DE PALABRAS MEDIANTE RELACIONES SINTAGMÁTICAS Y PARADIGMÁTICAS

T E S I S

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA

M. en C. JAVIER TEJADA CÁRCAMO

DIRECTOR: DR. ALEXANDER GELBUKH

DIRECTOR: DR. HIRAM CALVO

México, D.F.
Junio 2009

INSTITUTO POLITÉCNICO NACIONAL

SIP-14

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D. F. siendo las 15:00 horas del día 21 del mes de Noviembre de 2008 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis de grado titulada:

"CONSTRUCCIÓN AUTOMÁTICA DE UN MODELO DE ESPACIO DE PALABRAS MEDIANTE RELACIONES SINTAGMÁTICAS Y PARADIGMÁTICAS"

Presentada por el alumno:

TEJADA	CÁRCAMO	JAVIER LEANDRO
Apellido paterno	Materno	nombre(s)

Con registro: B 0 6 0 9 0 5

aspirante al grado de: ***DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN***

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***SU APROBACIÓN DE LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Presidente

Dr. Grigori Sidorov

Secretario

Dr. José Luis Oropeza Rodríguez

Primer vocal
(Director de Tesis)

Dr. Alexandre Felixovich Guelboukh Kahn

Segundo vocal
(Director de Tesis)

Dr. Francisco Hiram Calvo Castro

Tercer vocal

Dr. Héctor Jiménez Salazar

EL PRESIDENTE DEL COLEGIO

Dr. Jaime Álvarez Gallegos

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México, D.F., el día 13 del mes de mayo del año 2009, el que suscribe, Javier Leandro Tejada Cárcamo, alumno del programa de doctorado en ciencias de la computación, con número de registro B060905, adscrito al Centro de Investigación en Computación, manifiesta que el autor intelectual del presente trabajo de tesis bajo la dirección del Dr. Alexander Gelbukh Khan y del Dr. Hiram Calvo Castro, cede los derechos del trabajo intitulado: "Construcción automática de un modelo de espacio de palabras mediante relaciones sintagmáticas y paradigmáticas", al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección: jawitejada@hotmail.com. Si el permiso se otorga el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Javier Leandro Tejada Cárcamo

Resumen

Un modelo de espacio de palabras (WSM pos sus siglas en inglés: *Word Space Model*) es una representación espacial del significado de palabras, en la cual a cada vocablo se le asigna una localidad semántica tomando en cuenta sus propiedades de distribución en el lenguaje, de tal manera que la medición de la lejanía o cercanía entre dos vocablos determina su relación o similitud semántica.

En la actualidad, existe mucha investigación sobre diferentes técnicas de creación y recuperación automática de información de un WSM. Sin embargo, son pocos los estudios sobre el tipo de información usado en la construcción de este recurso. Existen factores que deben ser tomados en cuenta en la selección de un corpus: El tamaño, el origen del corpus (manual o automático) y el tipo de información (sintagmática o paradigmática).

Asimismo, evaluar el rendimiento de un WSM, no sólo depende del tipo de corpus usado en su construcción, sino del tipo de tarea de procesamiento de lenguaje natural a la que se aplica, tales como recuperación de información, traducción automática, minería de texto, respuesta automática a preguntas, etc. El impacto positivo o negativo de la información provista por un WSM en cada tarea es propio del método utilizado en su resolución, por lo que para valorar el verdadero impacto de la información que provee un WSM, se tendría que elegir una tarea en particular y un método de resolución específico.

En esta tesis se analiza el impacto de un WSM en la desambiguación de sentidos de palabras (WSD por sus siglas en inglés: *Word Sense Disambiguation*). Para esto se propone una arquitectura en la que un WSM proporciona *términos relacionados* sintagmáticamente o paradigmáticamente con una instancia ambigua (estos términos se seleccionan tomando en cuenta el contexto del vocablo ambiguo), los cuales son utilizados por un algoritmo de desambiguación. La combinación de dicho algoritmo con el WSM ha permitido superar los resultados reportados por el mejor método no supervisado orientado a la resolución de WSD que existe hasta el momento.

Abstract

A Word Space Model (WSM) is a spatial representation of the word meanings, in which to each word a semantic locality is assigned taking into account its distributional properties over the language, hence, the measurement of distance or proximity between two words determines their semantic relation or similarity.

Nowadays, there is much research on different techniques of automatic creation of WSMs and information retrieval using a WSM. Nevertheless, there are only few studies about the type of information used in building of this resource. There are factors that must be taken into account over the selection of a corpus: its size, its origin (manual or automatic), and the required type of information in the WSM (syntagmatic or paradigmatic).

Also, evaluation of the performance of a WSM depends not only on the kind of corpus used in its construction, but on the kind of natural language processing task to which it is applied, such as information retrieval, automatic translation, text mining, question answering, etc. The positive or negative impact of the information provided by a WSM in each task is due to the method used for the solution. This is why in order to evaluate the impact of the information provided by a WSM, we choose a particular task and a specific method of solution.

In this thesis we analyze the impact of a WSM to the word sense disambiguation (WSD) task. Due to this, we suggest an architecture in which a WSM provides terms related sintagmatically and paradigmatically with the current ambiguous instance (these terms are selected taking into account the context of the ambiguous word) under processing by a disambiguation algorithm. The combination of this algorithm with the WSM has improved the results that have been reported by the best unsupervised WSD method known in the literature.

Agradecimientos

Este trabajo no hubiera sido posible sin el apoyo de todo el personal del Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación. Muy en especial agradezco a mis directores: Dr. Alexander Gelbukh y Dr. Hiram Calvo Castro, y también a los doctores Grigori Sidorov e Igor Bolshakov.

Índice

Lista de tablas

Lista de gráficas

Lista de figuras

Capítulo 1
Introducción

El lenguaje natural es parte integral de nuestras vidas, siendo éste el principal vehículo usado por los seres humanos para comunicarse e intercambiar información. Tiene el potencial de expresar una gran cantidad de ideas; incluso elaborar y comprender pensamientos muy complejos. La lingüística computacional tiene por objetivo capturar este poder, suministrando la funcionalidad necesaria a computadoras para que éstas puedan analizar y procesar lenguaje natural, y de manera análoga, intenta comprender cómo las personas lo hacen.

Muchas aplicaciones de lenguaje natural, tales como traducción automática, desambiguación de sentidos de palabras, categorización de textos, creación automática de resúmenes, etc. se intentan resolver utilizando diversos métodos y técnicas empleando recursos construidos manual o automáticamente; sin conseguir aún una solución satisfactoria. Todas estas soluciones se basan en estadísticas matemáticas o diversas estructuras de organización del lenguaje las cuales intentan encontrar el *significado* expresado en el texto. Pero; ¿Qué es el *significado*? ¿Cuál es el significado de *significado*? La respuesta a esta pregunta aún permanece en las tinieblas. Quizás éste sea el "Santo Grial" de la lingüística computacional; y posiblemente el día que sea descubierto y definido cabalmente, las aplicaciones de lenguaje natural habrán sido solucionadas íntegramente.

1.1 Ubicación

Cuando se habla de semántica del lenguaje, éste suele representarse mediante diferentes estructuras de datos como árboles o grafos. Existe una estructura muy utilizada en la representación del texto, la cual permite determinar la lejanía o cercanía semántica entre un par de vocablos, tomando en cuenta su distribución con el resto de elementos del lenguaje. Dicha estructura, conocida como Modelo de espacio de palabras (WSM por sus siglas en ingles *Word Space Model*), determina la afinidad semántica entre dos vocablos usando un espacio multidimensional, cuyo número de dimensiones *n*, depende del número de vocablos diferentes encontrados en el corpus de texto usado

en la etapa de entrenamiento o construcción. Cada vocablo se representa mediante un vector de *n* dimensiones, que determina la distribución de éste con los demás elementos del sistema. Finalmente, la similitud semántica de dos palabras se cuantifica mediante el coseno del ángulo que forman sus vectores. De esta manera se seleccionan vocablos que comparten el mismo tópico o grupo semántico, los cuales son de gran utilidad en casi todas las áreas del procesamiento del lenguaje natural (NLP por sus siglas en inglés Natural Language Programming).

Tradicionalmente, la investigación realizada sobre WSM siempre ha estado enfocada hacia la creación de métodos para su construcción automática, así como diferentes técnicas para la explotación de la información que almacena este recurso. Algunos trabajos típicos en esta área son: LSA (por sus siglas en inglés *Latent Semantic Analysis*) [24], HAL (por sus siglas en inglés *Hyperspace Anlogue to Language*) [35], RI (por sus siglas en inglés *Random Indexing*), etc.

Paradójicamente, existen muy pocos trabajos dedicados al estudio de los corpus textuales, contenedores de información, que se usan en la construcción de este modelo. Existen tres características que se ha tomado en cuenta en este estudio:

- Origen del corpus. Hay corpus que son creados manualmente y otros automáticamente mediante algún programa. Por lo general, los corpus manuales son más pequeños que los automáticos. Por ejemplo, el corpus de Semcor es manual y el suministrado por Google ha sido creado automáticamente.

- Tamaño del corpus. Esta característica, en teoría, condiciona la información que proporciona un WSM. Se tiene la creencia que mientras más grande es el corpus, las estadísticas más confiables y por ende mejores resultados. Se han utilizado corpus de distintas dimensiones: British National de 100 millones de palabras, Tasa de 10 millones de palabras y Semcor de menos de un millón. Asimismo, se ha usado el corpus de Google cuyas especificaciones se detallan en el capítulo 5.

- Tipo de información. No todos los corpus proveen el mismo tipo de información. Los tesauros, por ejemplo, proveen vocablos con relaciones paradigmáticas (ejemplos de este tipo de relación son los sinónimos y antónimos). Los tesauros usados son el tesauro de Lin [32] y tesauro de Mobby. Otros corpus, como el de

Google, proporcionan vocablos relacionados sintagmáticamente (un ejemplo de este tipo de relación son las colocaciones gramaticales).

El impacto de un conjunto de palabras relacionadas semánticamente (proporcionadas por un WSM, un tesauro o cualquier otro recurso léxico), en las diferentes tareas que conciernen al procesamiento de lenguaje natural, no sólo depende de este conjunto de palabras, sino del proceso utilizado en la resolución de cada tarea. Sería largo y costoso evaluar tal impacto en varias de estas tareas. Se ha elegido la desambiguación de sentidos de palabras (WSD por sus siglas en inglés *Word Sense Disambiguation*) como tarea de evaluación, por ser una *tarea indirecta* en PLN (Procesamiento de Lenguaje Natural), es decir, que es utilizada por otras tareas más específicas como recuperación de información, creación automática de resúmenes, esteganografía, minería de texto, pregunta-respuesta (más conocido como *question-answering* en el idioma inglés), etc.

Con este propósito, se presenta una arquitectura en la que un WSM proporciona *términos relacionados* sintagmática o paradigmáticamente con una instancia ambigua (éstos se seleccionan tomando en cuenta el contexto del vocablo ambiguo), los cuales son utilizados por un algoritmo de desambiguación. Esta combinación ha permitido superar los resultados reportados por el mejor método no supervisado orientado a la resolución de la ambigüedad de sentidos que existe hasta el momento. [63]

1.2 Hipótesis

Es posible construir un modelo de espacio de palabras que proporcione vocablos relacionados sintagmática o paradigmáticamente, los cuales contribuyan en la solución satisfactoria de las aplicaciones de lenguaje natural, muy en particular de la desambiguación de sentidos de palabras.

1.3 Objetivos generales

– Construir modelos de espacios de palabras que proporcionen vocablos relacionados sintagmática y paradigmáticamente.

– Crear un método de desambiguación de sentidos de palabras cuyo éxito dependa del conjunto de términos relacionados semánticamente con la instancia ambigua, los cuales serán proporcionados por un WSM.

1.4 Objetivos específicos

– Determinar las características que influyen positiva o negativamente en la construcción de un modelo de espacio de palabras.

– Determinar el rol de los vocablos relacionados sintagmática y paradigmáticamente en la desambiguación de sentidos de palabras.

– Comparar los vocablos que proporciona un WSM con los existentes en otros recursos léxicos creados manualmente, como el tesauro de Mobby, o automáticamente, como el tesauro de Lin y el corpus de Google.

1.5 Motivación e importancia

Este trabajo ha sido motivado por el desarrollo de mi tesis de maestría: "Desambiguación de sentidos de palabras usando relaciones sintácticas como contexto local y medidas de similitud semántica". En esta tesis se presentó un método para desambiguar vocablos en forma automática, el cual utiliza algunos recursos léxicos y herramientas lingüísticas, tales como una base de datos de contextos sintácticos, WordNet [41] como repositorio semántico, *WordNet similarity package* [46] como herramienta que computa la relación semántica entre dos vocablos sobre WordNet, MINIPAR [33] como analizador morfológico y sintáctico para la extracción del contexto sintáctico de la instancia ambigua, etc.

En general, el método toma el *contexto sintáctico* del vocablo ambiguo para consultar una base de datos de *contextos sintácticos* con la finalidad de obtener términos que se usan en contextos similares. Este conjunto de vocablos, que comparten cierta afinidad semántica entre ellos, definen el sentido del vocablo polisémico mediante un algoritmo de maximización, en el cual cada *vecino* vota por un sentido; de tal manera

que el sentido con mayor número de votos es el elegido.

Intentar obtener *vecinos* correctos, es decir que se usen en contextos similares, es un proceso arduo y difícil. Para ello, primero se construyó un *recurso sintáctico* tomando un corpus extenso como origen de información (*British Nacional Corpus* de 100 millones de palabras). Dicho *recurso* recopila contextos sintácticos por cada palabra existente en el corpus. Luego, se compara el *contexto sintáctico* del vocablo ambiguo con los existentes en el recurso, ponderando la similitud de los contextos. Finalmente, las palabras correspondientes a tales contextos, forman el conjunto de *vecinos*.

Ahora bien, el éxito del algoritmo de maximización planteado depende de la calidad semántica de los *términos* que procesa y no de la cantidad de estos [63]. Por ejemplo, si se desea desambiguar el vocablo *estrella* en la siguiente oración:

Las estrellas del firmamento lucen más hermosas que nunca.

Si después de consultar al *recurso léxico* construido obtuviéramos los siguientes vecinos: *planeta*, *astro*, *universo*, *cometa*, *plutón*, *tierra*, entonces el algoritmo induce el sentido de *estrella* como *cuerpo celeste*. Si los *vecinos* obtenidos hubieran sido: *cantante*, *música*, *fama*, *dinero*, *glamour*, el algoritmo induce el sentido de estrella correspondiente a *persona famosa*. Este ejemplo puede ilustrarnos la importancia de la naturaleza semántica de los *vecinos* para inducir el sentido de un vocablo ambiguo; sin embargo la realidad no es tan cierta, ya que el recurso léxico proporciona *vecinos* de grupos semánticos diferentes, tales como *planeta*, *astro*, *universo*, *cometa, cantante*, *música*, *fama*, *dinero*. Esta deficiencia hace que el algoritmo falle.

Ante tales deficiencias surge la necesidad de construir un modelo de espacio de palabras (WSM) que proporcione *vecinos* de grupos semánticos diferentes. Un WSM es una arquitectura multi-vectorial de *n* dimensiones, en el cual cada vocablo se representa como un vector tomando en cuenta su distribución con los demás vocablos que conforman el sistema. Una hipótesis inicial considera que los grupos que se intenta obtener de un WSM son los que proporcionaría un tesauro creado manualmente, sin embargo según estudios realizados por Magnus Sahlgren [54], la intersección entre ambos recursos léxicos no supera el 10%. La ventaja de un WSM es, que a diferencia de los tesauros manuales, sólo toma en cuenta la información contenida en el texto,

mientras que los tesauros manuales están condicionados por las asociaciones semánticas que determinan los lexicógrafos que lo construyen.

1.6 Preguntas de investigación

A continuación se especifica una serie de preguntas de investigación que han sido respondidas con esta investigación.

- ¿Es necesario crear un WSM para obtener una *lista ponderada de términos relacionados* o es mejor usar otros recursos léxicos, tales como tesauros manuales o automáticos u otro tipo de corpus como el de n-gramas que provee Google?

- ¿Qué tipo de información debe de ser usada en la construcción de un WSM para maximizar la calidad semántica de los *términos relacionados* y por ende el algoritmo de desambiguación propuesto?

- Los *términos relacionados* que proporciona el WSM dependen del contexto de la instancia ambigua. ¿Qué tipo de contexto es el más recomendable para maximizar los resultados del método planteado?

- Entre los *terminos relacionados* que proporciona el WSM, existen *términos* más y menos relacionados con el vocablo polisémico. El algoritmo de desambiguación planteado utiliza estos *términos* para elegir un sentido para el vocablo ambiguo. ¿Cómo influye la calidad semántica de los vocablos en este proceso? ¿Cuánto términos son necesarios para desambiguar exitosamente una palabra?

1.7 Aportaciones

La principal aportación de esta tesis doctoral es la creación de un método para la construcción automática de un modelo de espacio de palabras, el cual proporciona términos relacionados sintagmáticamente y paradigmáticamente, los cuales mejoran el rendimiento del mejor método no supervisado que existe en la actualidad para la resolución de la ambigüedad de sentidos de palabras.

Asimismo, en el desarrollo de este trabajo se han creado recursos léxicos que pueden ser utilizados por otras tareas del procesamiento de lenguaje natural, tales como una base de datos de valores de similitud entre palabras y sentidos de palabras, y otra en la que se organiza la información del corpus de *n-gramas* de Google (36 gigabytes de información dispersos en 6 DVDs), para una acceso eficiente a dicha información. Ésta puede ser consultada en la dirección: 148.204.20.174/Google_corpus/default.aspx.

1.8 Publicaciones generadas

Hasta el momento se generaron las siguientes publicaciones; otras están en proceso de preparación.

- J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. Desambiguación de sentidos de palabras usando relaciones sintácticas como contexto local. Tutorials and Workshops of Fourth Mexican International Conference on Artificial Intelligence, ISBN 968-891-094-5, 2005. México.

- J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. Unsupervised WSD with a Dynamic Thesaurus. TSD 2007. 10th Internacional Conference o Text and Speech. República Checa.

- J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. An Innovative Two-Stage WSD Unsupervised Method. SEPLN. Revista 40, Marzo 2008. España.

- J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. The Role of Weighted Terms for Detecting the Most Frequent Sense in WSD. Enviado a revista ISI.

- J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. How to improve the semantic quality of a Word Space Model using different classes of information. Enviado a revista ISI.

Capítulo 2

Modelo de espacio de palabras

Schütze definió el modelo de espacio de palabras de la siguiente manera: “La similitud de vectores es la única información presente en este modelo, de tal manera que palabras que se relacionan semánticamente están *cerca* y aquellas que no se relacionan semánticamente están *lejos*”.[60]

WSM es una representación espacial del significado de palabras. Su idea fundamental radica en que la similitud semántica puede ser representada como proximidad o cercanía en un espacio de *n* dimensiones, donde *n* puede ser un número entero entre 1 y otro muy grande (puede llegar a millones de dimensiones). Ya que es imposible visualizar o imaginar dicho espacio, se puede ejemplificar usando un espacio de una o dos dimensiones (ver figuras 1 y 2).

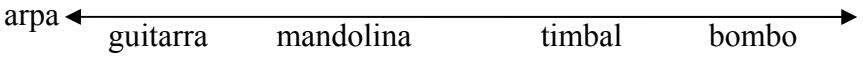


Figura 1 *WSM de una dimensión.*

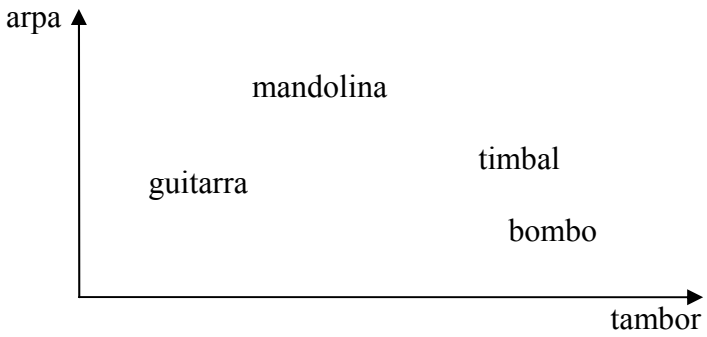


Figura 2 *WSM de dos dimensiones.*

En la figura 2, cada una de sus dimensiones representa un vocablo: El eje *x* al vocablo *tambor* y el eje *y* al vocablo *arpa*. Los demás vocablos (*guitarra*, *mandolina*, *timbal*, *bombo*) se van posicionando en el espacio dependiendo de la similitud que tienen con *arpa* y *tambor*. Se puede observar que las palabras tienden a formar grupos semánticos. Dicha agrupación, depende de la distribución que éstas presentan en el lenguaje; por ejemplo en un corpus de texto los vocablos que se usan con *timbal*, *bombo* y *tambor* suelen ser los mismos.

La factibilidad de representar una palabra en una arquitectura multidimensional tomando en cuenta su distribución en el lenguaje está muy demostrada. Es necesario tener en cuenta que la representación espacial de una palabra por si sólo no tiene sentido (ya que sólo sería un punto en el espacio multidimensional); sin embargo si otras palabras se representan en este espacio, es posible calcular similitudes y lejanías semánticas

2.1 La metáfora geométrica del significado

El uso de la proximidad espacial como una representación de la similitud semántica no es accidental ni mucho menos arbitraria; ya que para nosotros, los seres humanos, es natural conceptualizar similitudes usando nuestro conocimiento espacial y temporal del mundo. De esta manera conceptualizamos y discernimos el sentido de conceptos a veces muy abstractos. Esto ha sido sustentado por George Lakoff y Mark Johnson [23], quienes propone la *metáfora geométrica del significado*: "Los significados son localidades existentes en un espacio semántico, y la similitud semántica está denotada por la proximidad entre tales localidades".

Ellos intentan explicar la *similitud* como proximidad y *disimilitud* como lejanía. Esta metáfora, según Lakoff y Johnson, no es algo que nosotros tomamos y podemos dejar cuando queramos, sino es parte de nuestra existencia y la usamos para razonar. Por ende, esta afirmación no está basada en razonamiento intelectual sobre el lenguaje, sino lo establecen como un *prerrequisito* para razonar.

2.2 Hipótesis sobre la distribución del significado

Un WSM se construye automáticamente sin intervención humana y sin una base de conocimiento o reglas preestablecidas que restrinjan la similitud semántica entre palabras. Dichas similitudes son extraídas del texto siguiendo ciertos principios estructuralistas aplicados a este modelo por Schütze. [59]

La hipótesis sobre la distribución del significado afirma: "Palabras con distribuciones similares presentan significados similares". Un WSM usa estadísticas

para cuantificar las propiedades similares de distribución de las palabras. La idea consiste en colocar palabras en un espacio semántico tomando en cuenta su distribución en el texto con respecto a otras, de tal manera que la proximidad refleja similitud en la distribución.

Existen muchas investigaciones al respecto, tales como Schütze y Pedersen [59] quienes aseguran que palabras con significados similares ocurrirán con términos comunes y similares si se usa una cantidad de texto considerable. Rubenstein y Goodenought afirman que palabras con significados similares ocurren en contexto similares. [51]

Zellig Harris [15] desarrolló la *metodología de la distribución*. En su teoría afirma que es posible establecer entidades básicas del lenguaje, las cuales tienen cierta similitud en su distribución, de tal manera que sería posible conformar clases lingüísticas. Harris afirma que es posible establecer clases de tipos que engloben a todas las palabras del lenguaje. Para ello, sólo se debe tener en cuenta su distribución en el lenguaje y no necesita del concepto de *significado*. Es necesario mencionar que Harris estuvo fuertemente influenciado por Leonard Bloomfield, líder de la teoría americana del estructuralismo, aunque dichas definiciones fueron el legado del gran lingüista suizo Ferdinand de Sausurre (1857-1913).

2.3 Representación de un modelo de espacio de palabras

La manera clásica de representar un modelo de espacio de palabras es mediante una matriz. Los pioneros en la creación de algoritmos de implementación de esta matriz fueron Schütze y Qiu [59]. El algoritmo básico consiste en crear una matriz que almacene las frecuencias de las coocurrencias de palabras y los vectores de contexto se representan por las filas o las columnas de la matriz. Existen dos tipos de esta matriz: $w \times w$, (*word by word*) donde w es un vocablo (conocido como *word type* en inglés) y $w \times d$ (*word by* document) donde d es cada uno de los documentos que puede contener el origen de la información. Una celda f_{ij} de dicha matriz representa la frecuencia de ocurrencia entre la palabra i y otro vocablo o documento j. Las personas que trabajan en el área de recuperación de información (*information recovery*) usan las matrices del tipo

w x *d* como instancias del Modelo de Espacio Vectorial (MEV) desarrollado por Gerald Salton y sus colegas. [56]

Actualmente, existen diferentes esquemas de asignación de peso, tales como: el binario (1 si existe una relación entre dos palabras o entre una palabra y un documento y 0 en caso contrario), el esquema TF-IDF (por sus siglas en inglés *Term Frecuency - Inverse Document Frecuency*) que es uno de los más utilizados, el normal, donde sólo se toma la frecuencia de coocurrencia entre vocablos o entre vocablo y documento, son algunos de los más comunes.

2.3.1 ¿Qué clase de información proporciona un WSM?

Un WSM sólo proporciona palabras relacionadas y no tipifica su relación. Asimismo, la construcción del WSM sólo toma en cuenta la información existente en el texto. Ésta es la principal característica que diferencia un conjunto de palabras relacionadas semánticamente suministradas por un WSM con las que proporcionaría un ser humano, las cuales están influenciadas por otros factores, tales como la cultura social e intelectual de dicho individuo. Esto se demostró en los experimentos realizado por Magnus Sahlgren [55], en los que compara grupos de palabras proporcionadas por el WSM con los del tesauro manual Mobby, llegando a sólo 10% el valor de dicha intersección.

La pregunta ¿Qué clase de información se debe de usar en la construcción de un WSM para maximizar su rendimiento? no ha sido aún respondida, ya que existen muy pocos trabajos que aborden dicho tópico. Muy por el contrario, existe una gran variedad de investigaciones sobre otras características del modelo de espacio de palabras, tales como: diferentes esquemas estadísticos para ponderar la relación semántica entre un par de vocablos, medidas de similitud y lejanía semántica, reducción de la alta dimensionalidad que existe en el modelo, dispersión de los datos (*data sparseness*). Por ejemplo, Nakov *et al*. [42] estudió los diferentes esquemas de asignación de pesos a la coocurrencias de palabras en arquitecturas LSA (por sus siglas en inglés *Latent Semántic Analysis*), Bingham y Mannila [4], investigaron los efectos de usar diferentes técnicas de reducción de dimensionalidad y Weeds *et al*. [64] estudió diferentes medidas de similitud entre palabras. El estudio sobre los diferentes tipos de información

con los que se construye un modelo de espacio de palabras es importante por las siguientes razones:

- Las WSM basados en matrices *word by document* tienden a proveer vocablos relacionados sintagmáticamente, mientras que los basados en matrices *word by word* tienden a proveer información paradigmática. Determinar el impacto de ambos grupos de palabras en las diferentes tareas del procesamiento del lenguaje natural sería muy extenso; por ende acotamos dicho proceso orientándolo a la desambiguación de sentidos de palabras.

- Existen recursos léxicos que proveen información sintagmática, tales como el corpus de Google, e información paradigmática tales como el tesauro de Lin [34], el tesauro de Mobby, etc. Es necesario determinar si realmente conviene crear un WSM para obtener dicha información o quizás sea suficiente utilizar los recursos ya existentes.

2.3.2 Similitud semántica

La similitud semántica establecida en WSM ha desencadenado varias críticas al respecto. Padó y Lapata [43] afirman que dicho concepto es demasiado amplio como para ser útil, ya que existe relaciones semánticas diferentes tales como la sinonimia, antonimia, hiponimia, meronimia, entre otras. Espacios semánticos simplistas basados en WSM no pueden representar o distinguir tales relaciones. Estas críticas son válidas desde un punto de vista prescriptivo donde las relaciones deben de forma parte de una ontología lingüística. Desde un punto de vista descriptivo estas relaciones no son *axiomáticas* y la amplitud de similitud semántica representada por WSM es totalmente válida. Existen estudios que demuestran la realidad psicológica del concepto de similitud semántica. Por ejemplo, Miller y Charles [39] afirman que los seres humanos tienen la capacidad de inferir esta similitud casi instintivamente sin la necesidad de explicaciones más serias sobre dicho concepto.

Sin embargo, es válido afirmar que un WSM sólo representa palabras relacionadas semánticamente y no tipifica la relación (como sinonimia, antonimia, etc.); ya que para ello sería necesario modificar la hipótesis de distribución del significado.

2.4 Relaciones entre palabras

Existen dos tipos de relaciones entre palabras basadas en las propiedades de distribución de éstas en el lenguaje: Las relaciones sintagmáticas y las paradigmáticas. Las primeras se refieren a *posicionamiento*, específicamente palabras que coocurren en el texto. Ésta es una relación lineal que aplica a entidades lingüísticas que cumplen combinaciones secuenciales, como las de cualquier oración. Siguiendo este concepto una palabra puede ser combinada con cualquier otra. Un sintagma es una combinación ordenada de entidades lingüísticas. Por ejemplo, las palabras escritas son sintagmas de letras, las oraciones son sintagmas de palabras y los párrafos son sintagmas de oraciones.

Por otro lado las relaciones paradigmáticas se refieren a *sustitución*, es decir relacionan entidades que no coocurren en el texto. Estas relaciones se presentan entre unidades lingüísticas que ocurren en el mismo contexto pero no necesariamente juntas, como por ejemplo: *buenas noticias* y *malas noticias*, en cuyo caso las palabras *buenas* y *malas* tienen una relación paradigmática. Un paradigma es un conjunto de palabras que pueden sustituirse entre ellas. La tabla 1 ilustra claramente la diferencia entre ambas relaciones:

	Relaciones paradigmáticas (*x* OR z OR ...)		
Relaciones sintagmáticas (*x* AND *z* AND ...)	el	juega	basket
	ella	come	manzanas
	el	recoge	basura

Tabla 1 *Relaciones sintagmáticas y paradigmáticas.*

Capítulo 3

Desambiguación de sentidos de palabras

La ambigüedad surge en el lenguaje natural cuando una estructura gramatical puede ser interpretada de varias maneras. La desambiguación de sentidos de palabras (WSD por sus siglas en inglés *word sense desambiguation*), consiste en identificar el sentido de un vocablo ambiguo en un determinado contexto usando un conjunto de vocablos establecidos, por ejemplo, en inglés el vocablo *bat* podría ser un pequeño animal nocturno o una pieza de madera para hacer deporte y *bank* podría hacer referencia a un banco de peces o a una institución financiera. Cuando se usa un diccionario para encontrar la definición de una palabra, es posible verificar los diversos sentidos que ésta presenta, los cuales pueden ser totalmente diferentes, por ejemplo los sentidos del vocablo *consult* son *pedir un consejo* y *dar un consejo*.

3.1 Aplicaciones que requieren resolver la ambigüedad

La desambiguación no es un fin en sí misma, sino un proceso intermedio muy necesario para algunas tareas del procesamiento del lenguaje natural, tales como traducción automática, recuperación de información, extracción de información, categorización de textos, etc.

La traducción automática requiere al menos dos etapas: entender el significado de lo que se desea traducir, y una vez comprendido, generar las oraciones correctas al idioma destino. WSD es requerido en ambas etapas, ya que un vocablo puede tener más de una posible traducción en el lenguaje destino. Por ejemplo, la palabra inglesa *drug* puede ser traducida al turco como *ilac*, término que hace referencia al sentido de medicina o como *uyusturucu* para su correspondiente sentido de *dope* dependiendo del contexto donde ésta haya aparecido.

El área de recuperación de información también se beneficia de WSD, ya que la existencia de vocablos ambiguos en las consultas, es uno de los problemas principales en los sistemas de recuperación de información, los cuales necesitan módulos de WSD

para no mostrar aquellos documentos cuyos sentidos no son relevantes para la consulta.

Para los sistemas de procesamiento de habla, es importante determinar la correcta pronunciación de las palabras para generar sonidos naturales. Este proceso es muy difícil, ya que existen vocablos que toman una entonación diferente teniendo en cuenta el sentido que desean expresar. En el artículo escrito por Stevenson [61] se menciona un ejemplo al respecto, en el cual se afirma que la palabra *lead* es pronunciada de una manera distinta cuando se refiere al sentido *be in fron*t que cuando se refiere al sentido *type of metal*. WSD podría ayudar a identificar el correcto sentido de una palabra en el texto para generar una pronunciación correcta. El problema inverso podría ocurrir en el reconocimiento de palabras homófonas; es decir, vocablos que son diferentes textualmente; pero se pronuncian de la misma manera.

Para analizar el contenido y temática de un texto, es necesario considerar la distribución de las categorías predefinidas de las palabras, ya que existen vocablos que por su propia naturaleza expresan algún concepto, idea o tema específico. WSD es útil cuando se desea encontrar vocablos similares al sentido específico de una palabra.

3.1.1 Tipos de ambigüedad

Tradicionalmente, se distinguen tres tipos de ambigüedad: ambigüedad léxica, semántica, y estructural o sintáctica. La ambigüedad léxica se encarga de procesar aquellos vocablos que pueden pertenecer a diferentes categorías gramaticales, por ejemplo el vocablo *para* puede desempeñarse como preposición o como alguna conjugación del verbo *parar* e incluso del verbo *parir*.

La ambigüedad semántica procesa aquellos vocablos que tienen múltiples significados, por ejemplo *banco* puede significar *banco de peces*, *banco para tomar asiento* o *institución financiera*. Asimismo, es posible que una misma estructura sintáctica exprese diferentes significados, como la oración *todos los estudiantes de la escuela hablan dos lenguas*, la cual podría significar que cada estudiante habla dos lenguas o que en la escuela sólo se hablan dos lenguas determinadas. Otro ejemplo de este tipo de ambigüedad se presenta en la oración *vi a tu hermano volando hacia París*, la cual no deja claro si la persona que vio al hermano y el hermano iban en el mismo avión o en caso contrario el hermano vuela.

La ambigüedad sintáctica, también conocida como ambigüedad estructural procesa aquellas oraciones que pueden tener más de una estructura sintáctica. Por ejemplo, en la oración *Juan vio a un hombre con un telescopio*, es posible extraer al menos dos árboles de constituyentes, tal como se muestra en la figura 3.

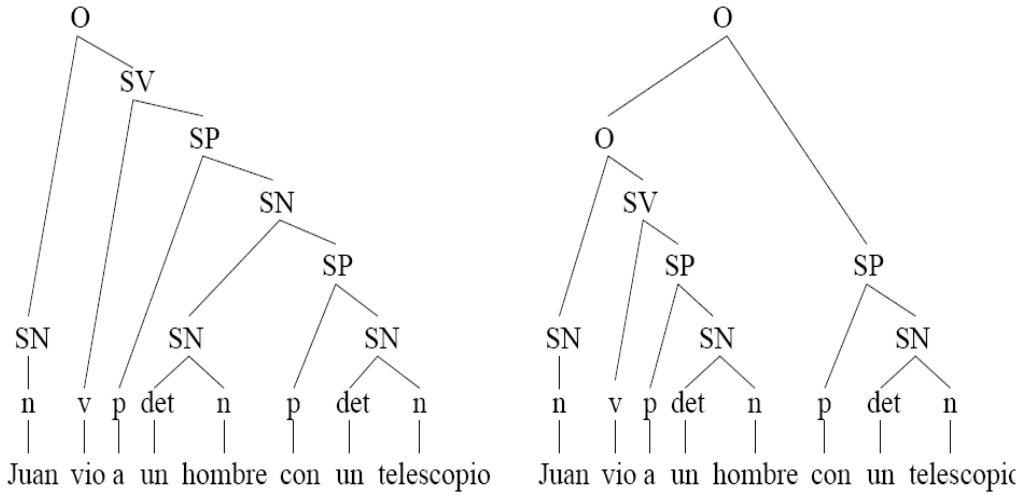


Figura 3 *Ejemplo de ambigüedad sintáctica en una oración*

3.1.2 Ambigüedad de sentidos de palabras

En la práctica resulta muy difícil distinguir el tipo de ambigüedad debido a la estrecha relación de dependencia entre los niveles clásicos del análisis lingüístico (morfológico, sintáctico, semántico y pragmático). Por ejemplo, el vocablo *traje* posee ambigüedad léxica, ya que puede ser verbo o sustantivo, y también ambigüedad semántica, ya que expresa diferentes ideas en cada caso.

Los trabajos realizados en esta tesis están orientados a solucionar el problema de la ambigüedad de sentidos de palabras o ambigüedad semántica a nivel de palabra. Esto significa etiquetar un vocablo ambiguo con el sentido que expresa en un contexto determinado. Es necesario tener en cuenta que el significado o idea específica de *sentido de palabra*, aún no se encuentra claramente definido por la comunidad científica, lo cual ha hecho que se desarrollen diferentes trabajos en WSD, cada uno tomando su propia versión acerca de dicha definición. Lo que si parece claro, es la posibilidad de distinguir entre desambiguación morfo-sintáctica y desambiguación de sentidos de palabras.

3.2 Resolución de ambigüedad de sentidos de palabras

Todos los trabajos de desambiguación de sentidos de palabras asocian el contexto (ya sea micro o macrocontexto) de un vocablo ambiguo con la información de un recurso de conocimiento externo (métodos basados en conocimiento), o con el contexto de instancias del mismo vocablo previamente desambiguado, las cuales son obtenidas de un corpus en una fase previa de entrenamiento (métodos basados en corpus). Cualquiera de estos métodos puede ser utilizado para asignar un sentido a cada ocurrencia de un vocablo ambiguo. La siguiente sección describe los recursos y técnicas empleadas por ambos métodos.

3.2.1 Métodos basados en conocimiento

En la década de los 70, diversas técnicas de inteligencia artificial eran usadas para desambiguar sentidos de palabras; sin embargo, esos trabajos se vieron limitados por la falta de recursos de información, surgiendo un gran problema, denominado *cuello de botella para la adquisición de conocimiento* [12]

Fue en los 80, cuando surgió una diversidad de materiales léxicos de gran escala, como diccionarios electrónicos, tesauros y corpus, los cuales dieron inicio a la extracción de información automática. Fue en esta época, que disminuyó el uso de teorías lingüísticas, las cuales fueron sustituidas por heurísticas para solucionar el problema de WSD. A continuación se explican los recursos más importantes que actualmente se usan por los métodos basados en conocimiento.

a. Uso de diccionarios electrónicos

Un diccionario electrónico surge de convertir un diccionario normal, creado exclusivamente para el uso humano, a formato electrónico. Éstos proveen información sobre sentidos de vocablos ambiguos, lo cual es explotado por el área de WSD.

El primer investigador que usó dichos diccionarios fue Lesk [30], quien creó una base de conocimiento, asociando a cada sentido una lista de palabras obtenidas de su definición. El proceso de desambiguación compara el contexto del vocablo ambiguo con la lista de cada uno de sus sentidos.

Otra variante, es comparar las listas de los vocablos del contexto con las del término ambiguo. Lesk logró una eficiencia de 50-70%; aunque el problema con este método es su *sensibilidad*; es decir, que la presencia o ausencia de algún vocablo contextual afecta radicalmente los resultados. Sin embargo, ésta técnica ha servido como base para muchos sistemas de WSD que usan diccionarios electrónicos como recurso de información.

Wilks [65] intentó mejorar el conocimiento asociado a la definición de cada sentido, calculando la frecuencia de coocurrencia entre los vocablos miembros de cada lista usando para ello un corpus de información. De esta manera, obtuvo diferentes medidas en cuanto al grado de relación semántica entre ellos. Finalmente, usó un método basado en vectores, el cual relaciona cada vocablo con su contexto y la métrica obtenida. Al experimentar con un solo término ambiguo, específicamente el vocablo *bank*, logró una eficiencia del 45% en etiquetado de sentidos y 90% en la detección de homógrafos.

Debido a que estos diccionarios fueron creados para personas y no para computadoras, se encuentran ciertas inconsistencias. Dichos recursos proporcionan información detallada a nivel léxico; sin embargo, omiten la información pragmática necesaria para determinar un sentido. Por ejemplo, la relación entre *ash* y *tobacco*, *cigarette* y *tray* no sería directa en una red semántica; sin embargo en el *Brown Corpus*, *ash* co-ocurre muy frecuentemente con ambos vocablos.

b. Uso de tesauros

El tesauro es un sistema que organiza el conocimiento basado en conceptos que muestran relaciones entre vocablos. Las relaciones expresadas comúnmente incluyen jerarquía, equivalencia y asociación (o relación). Los tesauros también proporcionan información como sinonimia, antonimia, homonimia, etc.

El tesauro de Roget denominado en ingles *Rogets´s Internacional Thesaurus*, fue convertido a versión electrónica en 1950, y ha sido el más usado en una amplia variedad de aplicaciones, tales como traducción automática, recuperación de información y análisis de contenido. Dicho tesauro suministra una jerarquía de conceptos de ocho niveles. Típicamente, la ocurrencia de un mismo vocablo bajo

diferentes categorías, representa un sentido diferente, de tal manera que un conjunto de términos asociados en una misma categoría se encuentran relacionados semánticamente.

La hipótesis básica para la desambiguación basada en tesauros, establece que cada una de las categorías semánticas de los vocablos que forman el contexto del término ambiguo, determinan la categoría semántica de todo el contexto. Finalmente, esta categoría es la que elige el sentido del vocablo ambiguo. Patrick [44] usó un tesauro para discriminar sentidos de verbos, examinando los grupos semánticos formados por cadenas derivadas del mismo. Dicho método es capaz de determinar el sentido correcto de verbos como *inspire* (*to raise the spirits vs. to inhale, breathe in, sniff*, etc.), *question* (*to doubt vs. to ask a question*) con un alto grado de confiabilidad.

Yarowsky [66] formó clases de palabras tomando como iniciales las definidas en las categorías del tesauro de Roget. Luego, para cada integrante de las clases formadas, obtuvo un contexto de cien palabras extraídas de la enciclopedia de Grolier, que en inglés se denomina *Grolier's Encyclopedia*, y usando técnicas estadísticas basadas en información mutua, identificó aquellas que coocurrían con los miembros de las clases formadas. Los grupos resultantes son usados para desambiguar nuevas ocurrencias de vocablos polisémicos, comparando una ventana contextual de cien vocablos con respecto al término ambiguo, con los miembros de cada grupo. Finalmente, usó la regla de Bayes para escoger un grupo, y como cada uno de ellos está enlazado a un sentido, etiquetó semánticamente la palabra polisémica. Este método reportó 92% de seguridad al realizar pruebas con vocablos ambiguos que expresan tres sentidos.

Al igual que los diccionarios electrónicos, un tesauro es un recurso creado por humanos, por consiguiente no contiene información correcta o real acerca de las relaciones entre palabras. Es ampliamente conocido que en los niveles más altos de la jerarquía de conceptos existe cierta contrariedad (aunque esto es cierto para cualquier jerarquía de conceptos), debido a que dichos conceptos son muy amplios como para poder establecer categorías semánticas.

Pese a estos problemas, los tesauros proporcionan una red muy rica en cuanto a asociaciones y relaciones entre palabras y un conjunto de categorías semánticas muy importantes para las diferentes tareas del procesamiento del lenguaje natural.

c. Uso de diccionarios orientados a la computación

Los diccionarios orientados a la computación, que en inglés se denominan *computacional lexicons*, son bases de conocimiento de gran escala, los cuales se empezaron a construir a mediados de los años 80. Algunos ejemplos son WordNet [40], CyC [29], ACQUILEX [5], COLMES [14], etc.

Existen dos técnicas fundamentales para la construcción de estos recursos: la técnica enumerativa y la técnica generativa. En la primera, los sentidos para cada vocablo ambiguo son proporcionados de manera explícita. En este grupo se encuentra WordNet, el cual se ha convertido en el diccionario computacional más usado para desambiguación de sentidos de palabras en inglés; sin embargo, presenta algunos defectos, tal como la especificidad de los sentidos definidos en WordNet; es decir, la manera tan perfecta como éstos han sido definidos, lo cual muchas veces no se ajustan a las necesidades requeridas por las aplicaciones de procesamiento de lenguaje natural y en especial WSD. Es necesario mencionar que aún no se tiene definido el nivel de especificidad necesaria para poder detectar las diferencias entre varios sentidos de un vocablo ambiguo, inclusive no es posible afirmar si las diferentes jerarquías definidas WordNet pueden o no satisfacer dichas necesidades. Actualmente, la comunidad científica para el procesamiento de lenguaje natural, está enfocando sus investigaciones a esta área.

Recientemente, algunos trabajos en WSD han utilizado diccionarios computacionales generativos. En estos recursos, las relaciones entre sentidos no son prescritas de manera explícita, sino son generadas por reglas que capturan las regularidades existentes al momento de crear las definiciones de dichos sentidos (como metonimia, meronimia, etc.). Buitelaar [7], especifica que la desambiguación de sentidos usando un contexto generativo, empieza con un etiquetado semántico, el cual apunta a una representación compleja del conocimiento, capturando los sentidos relacionados a un vocablo ambiguo de manera sistemática. Como segundo paso, el procesamiento semántico debe derivar interpretaciones dependientes del discurso, obteniendo información más precisa acerca del sentido de la ocurrencia dada. Buitelaar describe el uso de CORELEX, un diccionario para etiquetado semántico no especificado.

La gran limitante de estos recursos son sus dimensiones, ya que son más grandes que los descritos anteriormente. Buitelaar, describe una técnica que genera entradas automáticas para CORELEX empleando un corpus; así como los beneficios que se obtienen cuando se usan diccionarios de gran escala. Usando este método, es posible crear diccionarios orientados a un dominio específico.

3.2.2 Métodos basados en corpus

Un método basado en corpus explota un repositorio de ejemplos, a partir de los cuales se generan modelos matemáticos caracterizados por el uso de métodos empíricos. A mediados de los años 60, el uso de métodos estadísticos se vio menguado, debido al descubrimiento de reglas lingüísticas formales, tales como las teorías de Zellig Harris [15] y las teorías transformacionales de Noam Chomsky [8]. Por ende, los estudios se enfocaron al análisis lingüístico, tomando como referencia oraciones en vez del texto como un todo; así como el uso de ejemplos orientados a dominios específicos. Durante los siguientes quince años sólo un pequeño grupo de lingüistas siguieron trabajando con corpus, frecuentemente con fines lexicográficos y pedagógicos. Pese a ello, en esta época se crearon corpus importantes tales como: *Brown Corpus*, *Trésor de la Langue Française*, *Lancaster-Oslo-Bergen* (*LOB*) *Corpus*, etc.

a. Desambiguación supervisada

Los métodos de desambiguación supervisada etiquetan semánticamente un vocablo ambiguo tomando como referencia un repositorio de sentidos compilado previamente. Éste es entrenado con un corpus desambiguado (*training data*), donde para cada ocurrencia de una palabra ambigua, se toma el sentido de dicho vocablo y el contexto en el que se presenta. De esta manera, los algoritmos de aprendizaje pueden inferir, generalizar y aplicar reglas estadísticas e información lingüística tomando en cuenta dicho repositorio. Es necesario señalar, que los algoritmos de aprendizaje toman en cuenta los ejemplos del recurso como un conjunto de categorías, y que las personas que lo preparan han comprendido esta información, de tal manera que pueden combinarla con su propio conocimiento. En definitiva, el objetivo fundamental de la desambiguación supervisada es construir clasificadores capaces de diferenciar sentidos, basándose en el contexto adquirido previamente.

El principal problema de los métodos basados en aprendizaje supervisado, es la carencia de grandes corpus etiquetados semánticamente, los cuales no son suficientes para el entrenamiento de los clasificadores, así como las enormes cantidades de información generadas para cada sentido de un vocablo ambiguo. Pese a que los corpus etiquetados semánticamente de manera manual son extremadamente costosos, existen algunos recursos de este tipo, tales como: *The Linguistic Data Consortium* que proporciona aproximadamente 200,000 oraciones tomadas del *Brown Corpus*, el *Wall Street Journal*, en el que las ocurrencias de 191 palabras ambiguas son etiquetadas manualmente con los sentidos de WordNet, el *Cognitive Science Laboratory* proporciona 1,000 palabras tomadas del *Brown Corpus*, las cuales también han sido etiquetadas tomando como referencia WordNet.

Debido a la falta de estos recursos, se han realizado muchos trabajos para etiquetar automáticamente diversos corpus de entrenamiento, usando métodos basados en *bootstrapping*. En [16] se propuso un algoritmo, en cuya fase de entrenamiento se etiqueta semánticamente cada ocurrencia de un conjunto de sustantivos tomando en cuenta el contexto en el que aparece cada uno. Luego, la información estadística extraída del contexto es usada para desambiguar otras ocurrencias. Cuando algún vocablo es desambiguado con éxito, el sistema automáticamente adquiere información estadística adicional procedente del contexto de dicho término, mejorando el recurso de entrenamiento de manera incremental. Hearst indica que un conjunto de al menos diez ocurrencias son necesarias para iniciar el procedimiento de desambiguación y, para obtener una precisión alta son necesarias de veinte a treinta ocurrencias [16]. Otro método de *bootstrapping* basado en clases semánticas para dominios específicos ha sido propuesto por Basili. [3]

Brown [6] propuso el uso de corpus paralelos para evitar el uso de pequeños recursos etiquetados semánticamente. La idea es que diferentes sentidos de un vocablo polisémico, frecuentemente son traducidos de distintas maneras en otro lenguaje, por ejemplo *pen* en inglés es *stylo* en francés cuando expresa el sentido de escritura, y *enclos* cuando se refiere al sentido de envoltura. De esta manera al definir las ocurrencias de un vocablo ambiguo usando premisas de otro idioma, es posible determinar su sentido automáticamente.

Este método presenta algunas limitaciones, ente las cuales destacan las ambigüedades que son preservadas en el lenguaje destino (en francés *souris* y en inglés *mouse*) y la escasa disponibilidad de corpus paralelos de gran escala, incluso para el idioma inglés.

Uno de los principales problemas que presentan los métodos de desambiguación basados en corpus es conocido como *data sparseness*. Éste es causado por el uso de pequeños corpus de entrenamiento y por la diferencia de frecuencias que presentan los sentidos de un vocablo ambiguo en un corpus textual. Por ejemplo en el *Brown Corpus* (un millón de palabras), la palabra *ash* ocurre ocho veces, de las cuales sólo una de ellas hace referencia al sentido de *árbol*. Inclusive algunos sentidos de términos polisémicos no se encuentran en dicho corpus y para que un algoritmo de desambiguación sea exitoso, debe de asegurar que todos los sentidos de los vocablos polisémicos sean cubiertos.

b. Desambiguación no supervisada

La diferencia entre algoritmos de desambiguación supervisada y no supervisada, radica en que los primeros crean clasificadores usando ejemplos obtenidos de textos etiquetados semánticamente, los que a su vez han sido construidos de forma manual. Los métodos de desambiguación no supervisada usan la misma información para inferir características léxico-ambiguas en textos no etiquetados, obteniendo volúmenes de información más densos, solucionando parcialmente el problema de *data sparseness*, el cual parece estar estrechamente ligado con el uso de pequeños corpus etiquetados semánticamente. Dichos métodos son implementados usando modelos basados en clases y en similitud.

Los modelos basados en clases obtienen un conjunto de palabras que pertenecen a una categoría común, la cual es llamada clase. Brown, propone un método en el que dichas clases son derivadas de las propiedades de distribución inmersas en un corpus, mientras que otros autores usan información externa para definirlas. Resnik [48] usa las categorías de WordNet, Yarowsky [66] usa las categorías del tesauro de Roget, Lund [34] usa conjuntos conceptuales derivados de las definiciones de LDOCE (*Longman Dictionary of Contemporary English*).

Estos métodos confían en la premisa: *palabras de una misma clase comparten una temática común*, la cual es muy ambiciosa ya que la información contenida en dichas clases no siempre está relacionada con alguna temática específica. Por ejemplo, *residue* es un hiperónimo de *ash* en WordNet y sus hipónimos forman la clase {*ash*, *cotton*, *seed*, *cake, dottle*}. Obviamente los vocablos de este conjunto, se combinan de manera muy diferente en un texto, por ejemplo *volcano* está muy relacionado con *ash*; pero tiene poca o ninguna relación con los otros miembros del conjunto.

Los modelos basados en similitud explotan la misma idea; con la diferencia de que los vocablos no son agrupados en clases fijas, de tal manera que cada uno de ellos tiene un conjunto diferente de términos similares. Estos modelos explotan métricas entre patrones de coocurrencia. En [10] se experimentó con la pareja de vocablos (*chapter, describes*) los cuales pese a pertenecer a la misma clase, no aparecen juntos en el corpus usado; sin embargo los términos similares al vocablo *chapter*, tales como: *book*, *introduction*, *section* aparecen como pareja de *describes* en el mismo corpus. La evaluación de Dagan muestra que los métodos basados en similitud tienen mejor rendimiento que los basados en clases. McCarthy *et al*. [36] presentó un algoritmo que utilizando el tesauro de Lin, WordNet y ciertas medidas de similitud semántica, asigna el sentido más predominante a un vocablo ambiguo. En sus experimentos con sustantivos obtuvo 64% de aciertos.

Finalmente, el porcentaje de aciertos proporcionado por sistemas de desambiguación no supervisada es de 5% a 10% menor que los obtenidos con algoritmos supervisados.

3.2.3 El rol del contexto

El contexto es el indicador más relevante para identificar el sentido que expresa un vocablo ambiguo; por consiguiente, la mayoría de trabajos en WSD están enfocados a discriminar sentidos basados en él. Éste es usado de dos maneras:

- Como *bolsa de palabras*, la cual está conformada por un conjunto de términos alrededor del vocablo ambiguo. Para ello, se puede tomar en cuenta la distancia que existe entre los integrantes del contexto y el término ambiguo, de tal manera que las

relaciones gramaticales existentes en la oración son ignoradas.

- Como *información relacionada*, donde un conjunto de palabras son agrupadas tomando en cuenta algún tipo de relación o vínculo con el vocablo a desambiguar. Éstas pueden ser relaciones sintácticas, propiedades ortográficas, colocaciones gramaticales, categorías semánticas, preferencias de selección u otras.

El término contexto expresa un concepto muy amplio. Es por ello, que dependiendo del origen y la distancia de las palabras con respecto al término ambiguo, el contexto ha sido dividido en microcontexto, macrocontexto y dominio o temática específica.

a. Microcontexto

El contexto local o microcontexto está conformado por las palabras más cercanas al vocablo ambiguo, las cuales son seleccionadas teniendo en cuenta diversas características, tales como distancia, relaciones sintácticas y colocaciones gramaticales. Algunos trabajos basados en corpus usan microcontexto. Asimismo, ciertas técnicas basadas en diccionarios, usualmente no diferencian otro tipo de contexto que no sea el microcontexto.

Schütze [60] afirma que los métodos que tratan el contexto como *bolsa de palabras*, han logrado mejores resultados para sustantivos que para verbos; pero en general son menos efectivos que los métodos que toman en cuenta otro tipo de relaciones. Yarowsky [66], afirma que esta técnica es menos costosa que aquellas que requieren procesamientos más complejos y sus resultados pueden lograr niveles de desambiguación aceptables para determinadas aplicaciones.

Distancia

Esta característica hace referencia a un grupo de palabras cuya distancia con respecto al vocablo ambiguo no supera cierto umbral de referencia establecido. Los primeros trabajos realizados en WSD tomaban como contexto las palabras más cercanas al término a desambiguar. Este método que si bien es cierto, ha dado buenos resultados; aún dista mucho de ser lo suficientemente confiable.

Kaplan [19] afirma que existen pocos estudios que han intentado establecer la distancia óptima para lograr una desambiguación más confiable. Choueka y Lusignan [9] afirman que un contexto conformado por dos palabras es altamente confiable para lograr una desambiguación exitosa, incluso afirman que con una sola palabra es posible lograr un 80% de éxito.

Yarowsky [67] examinó el impacto de diversos umbrales tomando en cuenta palabras y duplas de palabras, ordenándolas dependiendo del acierto que tuvieron al desambiguar ocurrencias previas de términos ambiguos. Yarowsky llegó a la conclusión de que el umbral varía de acuerdo al tipo de ambigüedad. Afirma que las ambigüedades locales requieren un contexto de tres a cuatro palabras, mientras que las ambigüedades contextuales requieren de veinte a cincuenta palabras; sin embargo no se reportó una medida específica, ya que según sus experimentos cada vocablo ambiguo requiere diferentes umbrales y relaciones. Es necesario resaltar que Yarowsky usó como información adicional las categorías gramaticales de las palabras del contexto, lo cual hace difícil determinar el impacto del umbral en el proceso de desambiguación [68][69].

Colocaciones

Una colocación gramatical es un conjunto de dos o más palabras las cuales expresan una idea específica. El significado que expresa cada término de una colocación difiere de la semántica que dichos vocablos proporcionan cuando se usan de manera conjunta. El idioma inglés presenta muchas colocaciones; por ejemplo *cristal clear*, *middle management*, *nuclear family*, *cosmetic surger*, etc. Otro ejemplo más claro en cuanto al uso de colocaciones puede notarse en la expresión *red in the face*, la cual hace referencia a una persona apenada o sonrojada; sin embargo, *blue in the face* hace referencia a un individuo con hambre. Por ende, no sería común escuchar expresiones como *yellow in the face* o *green in the face*, ya que éstas denotarían posibles errores idiomáticos.

Algunos lingüistas, como Khellmer [21], argumentan que el *diccionario mental* del ser humano está conformado por colocaciones y vocablos individuales, de tal manera que un mismo vocablo relacionado con diferentes palabras expresa significados heterogéneos, como por ejemplo *bank river* y *bank investment.*

Kintsch y Mross [22] demostraron que el segundo término de una colocación define el sentido del primero. Yarowsky [67] fue uno de los primeros investigadores que usó colocaciones gramaticales para desambiguar sentidos de palabras. Él definió colocación gramatical como la coocurrencia de dos vocablos con alguna relación específica. Asimismo, concluyó que en el caso de ambigüedad binaria, existe un sentido por colocación, es decir, que el sentido que expresa un vocablo polisémico está determinado por el segundo término de la colocación a la que pertenece.

Relaciones sintácticas

En muchos trabajos de WSD, la información sintáctica es utilizada para determinar el sentido de una palabra ambigua. Wilks combina colocaciones gramaticales y relaciones sintácticas. Kelley, Dahlgren y Atkins especificaron ciertas reglas que justifican la presencia o ausencia de determinantes, pronombres, complementos de sustantivos, preposiciones, relaciones sujeto-verbo, verbo-objeto entre otras.

Hearst experimentó desambiguando sustantivos. Para ello, segmentó el texto en grupos de verbos, frases preposicionales y de sustantivos. Luego, examinó ítems que se encontraban aproximadamente a tres segmentos de frase del vocablo ambiguo. Yarowsky [67] observó varios tipos de comportamiento basados en la categoría sintáctica de las palabras y llegó a la conclusión de que el éxito de desambiguación de un vocablo puede depender de las estructuras sintácticas que lo rodean. De esta manera, afirma que los verbos derivan mejor información de sus objetos que de sus sujetos, que los adjetivos derivan mayor información de los sustantivos a los que modifican, y que los sustantivos son mejor desambiguados cuando tienen adjetivos o sustantivos adyacentes.

En trabajos recientes, la información sintáctica sólo es usada para categorizar gramaticalmente las palabras en conjunción con otra clase de información. La evidencia sugiere que la desambiguación exitosa de un vocablo depende del método usado, el cual debe de elegirse tomando en cuenta su categoría gramatical y su contexto local.

b. Macrocontexto

El macrocontexto, que en inglés se denomina *topical context*, está conformado por palabras de gran contenido semántico (por lo general sustantivos, adjetivos y verbos), las cuales coocurren con un sentido específico del vocablo ambiguo, usando varias oraciones como fuente de información.

A diferencia del microcontexto, el cual ha sido muy usado desde mediados del siglo pasado, el macrocontexto ha sido menos utilizado.

Los métodos que hacen uso de esta característica, explotan la redundancia en el texto; es decir, intentan ubicar grupos de palabras que se encuentren semánticamente relacionadas con un tópico específico. Por ejemplo, sería posible determinar el sentido *base-ball game* al que hace referencia el vocablo ambiguo *base*, si su macrocontexto está conformado por términos como *pitcher* y *ball*.

Por lo general, los trabajos basados en macrocontexto, generalmente usan una *bolsa de palabras* que agrupa un conjunto no ordenando de términos con gran contenido semántico, los cuales son encontrados en el macrocontexto del término ambiguo.

Yarowsky [66] usó un contexto de 100 palabras para obtener clases de términos relacionados semánticamente. Asimismo, usó dicha cantidad para desambiguar un vocablo polisémico tomando como referencia el tesauro de Roget. Gale [12] tomó un contexto de aproximadamente cincuenta palabras, indicando que mientras éstas son más cercanas al vocablo ambiguo, la elección del sentido es más confiable. Los resultados obtenidos por Gale sólo mejoraron de 86% a 90% cuando amplió el número de palabras en el contexto, de 6 (cantidad de términos que generalmente se usan como microcontexto) a 50 palabras.

En un estudio similar, Gale llegó a la conclusión que si en un discurso se presentan diferentes ocurrencias de un vocablo ambiguo, es muy probable que éstas hagan referencia al mismo sentido. Leacock [27][28] demostró que ambos tipos de contexto (microcontexto y *topical context*) son necesarios para obtener resultados más confiables. Los estudios de Yarowsky indican que la información obtenida del macrocontexto puede ser usada para desambiguar sustantivos, mientras que para verbos y adjetivos los resultados disminuyen dramáticamente cuando se usan muchas oraciones

como parte del contexto [68].

Brown y Yule [6] sugieren que los métodos que utilizan contextos muy amplios, deberían de dividir el texto en sub-tópicos, los que a su vez deberían de agruparse en secciones de textos que se encuentren conformados por diferentes párrafos. Ellos afirman que la segmentación automática de texto en unidades más pequeñas, podría ser de mucha ayuda para dichos métodos. Leacock [28] consideró el rol del micro y macrocontexto, intentando definir el impacto de cada uno de ellos. Sus resultados indican que para clasificadores estadísticos, el microcontexto es superior al macrocontexto.

c. Dominio

El término dominio en lenguaje natural puede ser definido como una agrupación de diferentes fuentes de información textuales que hagan referencia a un tópico específico de información. El hecho de desambiguar sentidos usando un dominio, se encuentra implícito en varias técnicas de WSD basadas en inteligencia artificial, como aquellas en las que se elige un sentido tomando en cuenta no sólo el contexto del término ambiguo; sino el sentido general del tópico o discurso. Los métodos basados en la hipótesis: una palabra sólo tiene un sentido por discurso, demuestran sus limitaciones cuando sólo usan esta fuente de información para desambiguar sentidos. Por ejemplo en la oración *the lawyer stopped at the bar for a drink*, si sólo se confiara en la información del tópico o dominio, el vocablo *bar* asumiría un sentido incorrecto, ya que para este caso el tópico hace referencia a leyes.

Gale cuestiona en demasía el hecho que una palabra sólo tenga un sentido por discurso. Dahlgren afirma que el dominio no elimina la ambigüedad para algunas palabras. Para ello, toma como ejemplo el sustantivo *hand* el cual tiene aproximadamente 16 sentidos y puede expresar 10 de ellos en cualquier texto.

La influencia del dominio para etiquetar semánticamente una palabra, depende de factores como el estilo del texto (cuan técnico es por ejemplo) y la relación existente entre los sentidos del vocablo ambiguo; es decir, si éstos están fuerte o débilmente polarizados, así como su uso especializado y su uso común. Por ejemplo, en la enciclopedia francesa *Encyclopaedia Universalis*, el término *intérêt* (interés) aparece 62

veces en el artículo *Interés-Finanzas*, en el que hace referencia al sentido de finanzas, 139 veces en el artículo *Interés-Filosofía*, en el que hace referencia a un sentido no financiero; sin embargo en el artículo *Tercer mundo*, aparece dos veces para cada uno de estos sentidos.

Capítulo 4

Método propuesto

Existen pocos trabajos que utilizan la información producida por un WSM en la resolución de la ambigüedad de sentidos palabras. El más conocido, propuesto por Gallan y Schütze [11][60] se limita a la creación de vectores de sentidos en un espacio de *n* dimensiones. En dicho trabajo, se tiene un vector predeterminado por cada uno de los posibles sentidos de la instancia ambigua, y luego se compara cada uno de ellos con el vector representante del vocablo ambiguo cuyas dimensiones están conformadas por su contexto. De esta manera, el vector predeterminado que matemáticamente sea más similar al vector de la palabra polisémica, es el que determinará su sentido.

En esta tesis doctoral, se propone una arquitectura en la que un WSM proporciona un conjunto de *términos relacionados* sintagmáticamente o paradigmáticamente con una instancia ambigua (estos se seleccionan tomando en cuenta el contexto del vocablo ambiguo), los cuales son utilizados por un algoritmo de desambiguación. Esta combinación ha permitido superar los resultados reportados por el mejor método no supervisado orientado a la resolución de la ambigüedad de sentidos que existe hasta el momento [63].

El método propuesto presenta tres etapas, que se pueden apreciar en la figura 4.

- En la primera, se crea automáticamente un modelo de espacio de palabras.
- En la segunda, se obtiene una *lista ponderada* de *términos relacionados* semánticamente con el vocablo ambiguo, los cuales se seleccionan tomando en cuenta su contexto. Esta lista se obtiene del WSM creado.
- En la tercera, se ejecuta un algoritmo de desambiguación el cual toma en cuenta la *lista ponderada* para determinar el sentido de la instancia ambigua

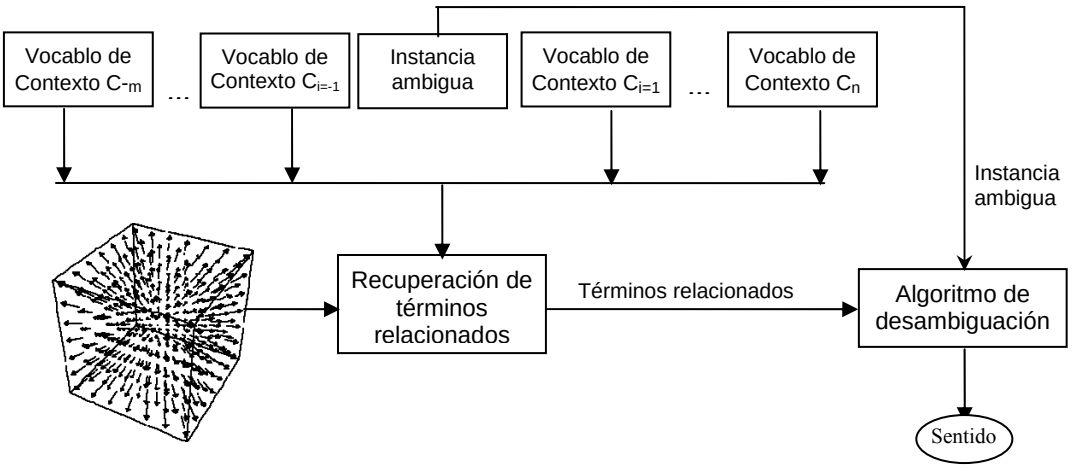


Figura 4 *Arquitectura planteada*

4.1 Construcción del modelo de espacio de palabras

La construcción automática de un WSM es un proceso que toma como entrada un corpus de texto y crea un espacio multidimensional en la cual se representa todos los vocablos existentes en el corpus, tomando en cuenta sus propiedades de distribución en el mismo.

El proceso que se realiza sobre el corpus depende del tipo de WSM que se desee crear: paradigmático o sintagmático (ver figura 5).

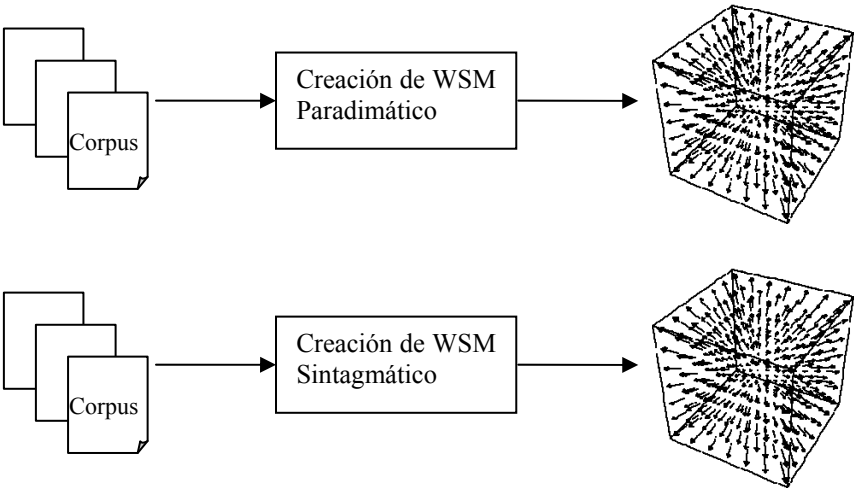


Figura 5 *Creación de un WSM*

4.1.1 Procesamiento del corpus

En esta sección se explica el tratamiento del corpus de entrenamiento, el cual varía dependiendo del tipo de modelo de espacio de palabras que se desea construir. La diferencia entre WSM sintagmático y paradigmático radica en el conjunto de *términos relacionados* que proporcionan.

Por ejemplo, si consultáramos a un WSM paradigmático los términos más similares al vocablo *universidad*, serían *academia*, *escuela*, *colegio*, etc., mientras que un WSM sintagmático hubiera proporcionado vocablos como: *privada*, *nacional*, *tecnológica*, etc.

a. Tratamiento paradigmático del contexto

Dos vocablos relacionados paradigmáticamente no coocurren entre ellos; sin embargo, los vocablos con los que coocurren suelen ser los mismos. Por ejemplo, adjetivos diferentes que modifican al mismo sustantivo, como *buena noticia* y *mala noticia*, o *querido padre* y *querida madre*. De dichas coocurrencias se puede determinar que los vocablos *malos y buenos*, *padre* y *madre* presentan una relación paradigmática o de sustitución.

Existen tres parámetros que influyen en la obtención de este tipo de relaciones: el tamaño de la ventana contextual, la posición de las palabras dentro de dicha ventana y la dirección en la que la región de contexto se extiende (hacia delante o atrás). Para ilustrar mejor este punto imaginemos dos secuencias de palabras:

bla bla bla ***blo*** bli bli bli
bla bla bla ***ble*** bli bli bli

Es fácil notar que las palabras *blo* y *ble* presentan una relación paradigmática ya que las palabras anteriores y posteriores de ambas son las mismas. En este caso no importa si tomamos una ventana de tamaño 1+1 (un vocablo a la izquierda y uno a la derecha), 2+2 e incluso 3+3, ya que en este caso particular, cada ventana confirma la relación paradigmática entre *blo* y *ble*.

Las ventanas contextuales pueden ser estáticas y dinámicas, es decir, con un

número de vocablos fijos o variables a la derecha e izquierda de la palabra en cuestión. Actualmente los investigadores prefieren ventanas estáticas.

Ahora supongamos el siguiente contexto:

bla ***blo*** e bli → ***blo***: (bla) + (0 bli) → ***blo***: (1) + (0 1)
bla ***ble*** h bli → ***ble***: (bla) + (0 bli) → ***ble***: (1) + (0 1)

Quizás cueste un poco más darse cuenta que las palabras *blo* y *ble* presentan una relación paradigmática, ya que los vocablos de la derecha (*e* y *h*) de cada una no son las mismas. Esto se representa mediante notación binaria:

blo: (1) + (0 1)
blo: (1) + (0 1)

Existen diferentes maneras de asignar peso a los vocablos de una ventana, pero las más comunes son la binaria y la asignación de mayor peso a aquellos que se encuentran más cercanos a la palabra en cuestión.

b. Tratamiento sintagmático del contexto

El caso más claro de vocablos relacionados sintagmáticamente son las *colocaciones gramaticales*, por ejemplo *prestar atención*, *tomar asiento*, *colgar los tenis*, etc. Una colocación está conformada por dos o más palabras que *trabajan* juntas para expresar una idea y pese a que pueden combinarse con otras, su correlación es muy común ya que suelen ocurrir una al lado de la otra sin la existencia de palabras intermedias. La característica principal es que el significado que expresan juntas es diferente al que cada una expresa por separado.

Asimismo, existen pares de vocablos que pese a estar separadas por otros (que pueden ser muchos) mantienen una relación sintagmática. Por ejemplo, el primer vocablo del primer párrafo de esta tesis tiene una relación sintagmática con el último vocablo del mismo párrafo. Por ende, el principal parámetro que influye en la obtención de vocablos relacionados sintagmáticamente es el tamaño de la *región contextual* en el que las ocurrencias se cuentan. Esta región puede ser una pequeña secuencia de palabras, como una oración o un párrafo, o una mayor, como la totalidad de un texto.

En los sistemas de recuperación de información, un documento es el contexto natural de una palabra. Dichos sistemas asumen a los documentos como *unidades tópicas* y sus palabras como *indicadores tópicos* cuya distribución es gobernada por un número limitado de *tópicos* (afirmación algo ambiciosa en nuestros días ya que un texto podría expresar muchos tópicos). Ahora bien, en un texto plano o un corpus inmenso existe una variedad de tópicos que se traslapan, de tal manera que el contexto más apropiado de una palabra es la *oración* en la que se presenta. Quizás este sea el contexto más usado por investigadores en el área.

Otra posibilidad, es usar una región contextual más reducida, en la que se busquen parejas de vocablos consecutivos, sin embargo habría que tomar en cuenta, que cuando deseamos obtener información sintagmática es necesario considerar que sólo muy pocas combinaciones de palabras (quizás sólo las colocaciones) coocurren frecuentemente en un pequeño contexto. Picard [47] afirmó que la mayoría de estos términos nunca coocurren, lo cual trae como consecuencia una estadística muy pobre y el problema de *sparse data*.

4.1.2 Construcción de la matriz

Ambos tipos de WSM son representados mediante una matriz, en la cual una fila representa a un vocablo existente en el corpus de entrenamiento, mientras que el número de columnas varía dependiendo del tipo de WSM que se construye:

- En el caso de los paradigmáticos se crea una columna por cada vocablo existente en el corpus, por ende una matriz paradigmática tiene dimensiones $n \times n$.

- En el caso de los sintagmáticos, el corpus de texto se divide en *regiones contextuales* del mismo tamaño (de 10 vocablos o más grandes, como 150). El número de columnas de esta matriz depende de la cantidad de *regiones contextuales* d en la que se divide el corpus de entrenamiento, por ende una matriz sintagmática tienen dimensiones $n \times d$.

Por ejemplo, dada la oración: *el gato se comió al ratón*. La matriz paradigmática resultante tomando en cuenta dicha frase y además, una ventana contextual de un solo vocablo a la izquierda y otro a la derecha se puede apreciar en la tabla 2.

Vocablo	gato	comer	ratón
gato	0	1	0
comer	1	0	1
ratón	0	0	1

Tabla 2 *Matriz paradigmática.*

El valor que se establece en las celdas de la matriz es un peso de ponderación, que se asigna tomando en cuenta diferentes esquemas estadísticos. En el ejemplo se usa un esquema binario, donde 1 significa que existe una coocurrencia entre la fila *i* y la columna *j* y 0 que dicha coocurrencia no existe.

En el método propuesto, sólo se toman en cuenta vocablos de contenido (*content words* en inglés), tales como sustantivos, adjetivos y verbos, desechando las palabras de parada (*stop words*), lo cual permite reducir las dimensiones de la matriz. Asimismo, se usan los *lemas* de los vocablos seleccionados.

4.1.3 Esquema de ponderación

El esquema de ponderación hace referencia a un valor que denota la afinidad o relación semántica entre la fila *i* y la columna *j*. En nuestro método este valor inicialmente es la frecuencia de correlación entre dos vocablos en el corpus (en el caso de los paradigmáticos) o la cantidad de veces que sucede un vocablo en una región contextual (en el caso de los sintagmáticos).

En este método se utiliza otro tipo de ponderación, conocido como el esquema TF-IDF (por sus siglas en inglés *Term Frecuency - Inverse Document Frecuency*), el cual generalmente se aplica a tareas de clasificación y similitud de documentos. En este esquema, cada documento es representado por un vector cuyo número de dimensiones corresponde a la cantidad de *vocablos* que existen en él.

En nuestro método, el valor en cada celda de la matriz está determinado por un peso *w* (ver ecuación 3), el cual se calcula como el producto del TF (ver ecuación 1) e IDF (ver ecuación 2). El peso $w_{(i,j)}$ determina el grado de relación semántica entre el *vocablo i* (la fila) y palabra *j* o sección *j* (la columna). El TF muestra la importancia de un *vocablo* respecto a la palabra que modifica o a la sección en la que se encuentra.

Por la tanto, el peso de la relación aumenta si el *vocablo* aparece más a menudo con dicha palabra o sección. El IDF denota la importancia de un *vocablo* respecto al resto de palabras del corpus, de tal manera que su peso disminuye si aparece más a menudo con los demás *vocablos* o *secciones* del corpus, y aumenta cuando aparece con la menor cantidad de estos, ya que los *vocablos* o *secciones* muy frecuentes discriminan poco a la hora de representar al *vocablo* mediante un vector.

$$f_{i,j} = \frac{freq_{i,j}}{\max freq_{l,j}} \quad (1)$$

$$idf_i = \log \frac{N}{n_i} \quad (2)$$

$$w_i = f_{i,j} \times idf_i \quad (3)$$

En las ecuaciones anteriores, *i* representa a la i-ésima fila (vocablos) y *j* (puede ser vocablos o secciones) representa a las j-ésima columna de nuestra matriz. La $freq_{i,j}$ es la frecuencia entre *i* y *j*, $max\ freq_{l,j}$ es la más alta de cualquier vocablo *i* en una sección o vocablo *j*. *N* es el número de vocablos del corpus tomados en cuenta en la construcción de la matriz, n_i es el número de vocablos con los que ocurre *j*, y w_i es el peso final. El peso *w* que se calcula para todas las dimensiones de un vocablo *i*, forman un vector que se almacena en el WSM (ver ecuación 4)

$$\vec{V}(vocablo_i) = \{(dim_1, w_1), ..., (dim_n, w_n)\} \quad (4)$$

$\vec{V}$ *(vocablo$_i$)* es el vector que representa al vocablo *i* con respecto a la totalidad de vocablos o secciones del corpus, dim_n cada una de las dimensiones del WSM (el número de dimensiones es el número de vocablos o secciones del corpus de entrenamiento) y w_n es el peso asignado a dim_n. Muchos pesos son 0 (cero), lo que denota la inexistencia de una correlación entre el vocablo *i* y el vocablo o sección *j*. Realmente la cantidad de ceros en el sistema es elevada (Data Sparseness), lo que confirma la Ley de Zipf [70], que indica que sólo unas pocas palabras en el lenguaje (*non content words*) se comportan *promiscuamente*, es decir se relacionan con muchas palabras.

Tomando en cuenta el ejemplo anterior, se puede determinar la lista de

coocurrencias de un vocablo; por ejemplo, en la oración mostrada en la tabla 2, el vocablo *comer*, está representado por la tupla (1,0,1), la cual se convierte en un vector $\vec{V} = (x_1, x_2, x_3)$ donde el número de dimensiones es el número de vocablos en el sistema (en nuestro ejemplo son tres), y los valores (1,0,1) son las coordenadas de posicionamiento en el espacio vectorial. Representando vectorialmente las propiedades de distribución de una palabra podemos pasar a una representación geométrica de dicho vocablo.

4.1.4 Recuperación de términos relacionados

En esta etapa se obtienen un conjunto de términos que se relacionan semánticamente con el vocablo ambiguo. Asumimos que dos vocablos mantienen esta relación cuando ambos se utilizan en contextos similares (en inglés se conoce como *second order context*). Como se ha explicado en esta sección, el WSM almacena vectores de dimensionalidad *n*, y cada uno de estos, debe de ser comparado con un vector que represente al vocablo ambiguo.

La dimensionalidad del vector ambiguo sigue siendo *n*, sin embargo el peso de ponderación en cada dimensión, depende de los vocablos de su contexto. El tipo de contexto es una variable que condiciona el poder semántico de este vector. En la tabla 3, las columnas d_i son las dimensiones de los vectores, y las filas son los vocablos del sistema. *w* es el vocablo ambiguo y su vector indica que los vocablos de su contexto son d_2 y d_4. La columna *similitud*, ordenada descendentemente, indica que el vocablo r_1 es más similar a *w*, que el vocablo r_n.

	d_1	d_2	d_3	d_4	d_5	d_6	…	d_n	Similitud
w	0	1	0	1	0	0	…	0	
r_1	1	2	6	5	2	2	…	0	0.99
r_2	0	5	4	0	1	1	…	3	0.92
r_3	4	4	3	3	1	1	…	0	0.83
…								…	
r_n	0	4	0	5	2	2	…	4	0.68
…								…	

Tabla 3 *Vectores en un WSM.*

Existen varias maneras de computar la similitud de dos vectores y en nuestro método se determina mediante el valor del coseno del ángulo que forman expresado por

el cociente entre el producto *punto* y el producto *cruz* (ver ecuación 5)

$$cos\_measure(\vec{w},\vec{r}_i) = \frac{\vec{w}\cdot\vec{r}_i}{\left|\vec{w}\right| \times \left|\vec{r}_i\right|} = \frac{\sum_{j=1}^{n} w_j \times r_{i,j}}{\sqrt{\sum_{j=1}^{n}(w_j)^2} \times \sqrt{\sum_{j=1}^{n}(r_{i,j})^2}} \quad (5)$$

Al seleccionar y ponderar los vectores más similares, se obtienen los vocablos más relacionados y, finalmente las *listas ponderadas*, que se representan mediante la ecuación 6. Cada par ordenado (r_n,w_n) representa a un *vocablo relacionado* y su peso de relación semántica con Voc_n

$$\text{Lista}(\text{Voc}_n) = \{(r_1, w_1),(r_2, w_2),...,(r_n, w_n)\} \quad (6)$$

4.2 Algoritmo de desambiguación

El algoritmo de desambiguación planteado, está basado en el propuesto por Diana McCarthy *et al*. [36]. En el nuestro, cada *vocablo* del conjunto de términos, emite un voto para alguno de los sentidos del vocablo ambiguo, de tal manera que el sentido que obtenga más votos, es el elegido. Esta decisión se basa en dos factores:

- El conjunto de términos utilizados. El método de McCarthy *et al*. consulta al tesauro de Lin para obtener los *vocablos relacionados* con el vocablo ambiguo. Nuestro método usa el WSM construido.

- La manera de cuantificar la relación semántica entre dos vocablos. Lo vocablos que provee el tesauro de Lin no se encuentran ponderados (no indica cuál de ellos, es el más y menos relacionado), por ende McCarthy *et al*. determina esta relación empleando diversas medidas de similitud semántica implementadas sobre WordNet. En nuestro método el WSM provee esta cuantificación.

El proceso en el cual un *vocablo relacionado* vota por un sentido es común en ambos métodos. Para ello, es necesario determinar la afinidad semántica entre cada uno de los sentidos del *vocablo relacionado* y del ambiguo. En la figura 6, *r*, tiene tres sentidos: r_1, r_2 y r_3 y w tiene dos: w_1 y w_2.

Una línea indica una comparación de sentidos. La línea contínua indica que los sentidos r_1 y w_2 son los más afines, lo cual indica que r vota por el sentido w_2 del vocablo w.

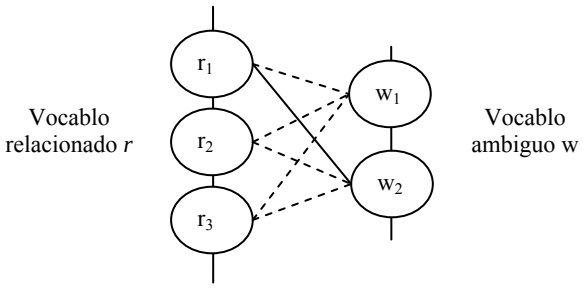


Figura 6 *Comparación de sentidos*

Un sentido en WordNet se representa mediante una glosa (una expresión que define el sentido). Para computar la afinidad semántica entre dos glosas se usa WordNet::Similarity[46], el cual es un conjunto de librerías que implementan medidas de similitud y relación semántica sobre WordNet, tales como las propuestas por Resnik[48], Lin [34], Jiang–Conrath [18], Leacock–Chodorow [26], entre otras.

Es importante aclarar que el método de McCarthy *et al*. está orientado a obtener el sentido predominante de un vocablo ambiguo, mientras que el nuestro determina el sentido que expresa una instancia ambigua en un contexto específico. Por esta razón, al momento de seleccionar los *vocablos relacionados*, nosotros usamos el contexto de la instancia ambigua en el proceso de selección, mientras que el método de McCarthy *et al*. no lo hace. A continuación se define la notación que se usa para explicar este algoritmo. Sea:

w, la palabra ambigua.

$S(w) = \{s_1, s_2, \ldots, s_i\}$ el conjunto de sentidos del vocablo ambiguo.

$R(w) = \{(r_1,w_1), (r_2,w_2), \ldots, (r_j,w_j)\}$, un conjunto de pares ordenados, donde cada (r_i,w_i), determina el peso w_i de la relación semántica entre r_i y w_i.

$S(r_j) = \{ r_{j1}, r_{j2}, \ldots, r_{jk} \}$ es el conjunto de sentidos de la palabra relacionada r_i.

El objetivo del algoritmo es asignar un puntaje a cada uno de los sentidos w_i. De

tal manera que el sentido que obtenga mayor puntaje es el elegido para el vocablo *w*.

El peso de s_i es la sumatoria del peso de cada una de sus palabras relacionadas (r_i,w_i) multiplicada por un valor. Este valor es la máxima similitud semántica entre el sentido s_i y cada uno de los sentidos $S(r_i)$, dividido por la sumatoria de la máxima similitud semántica entre cada sentido S(s) y los sentidos $S(r_i)$. La figura 7 y la ecuación 7 ilustran claramente este proceso.

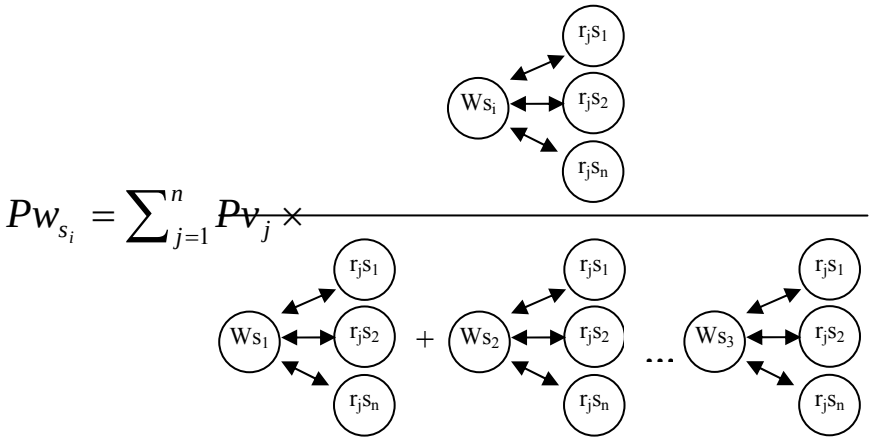


Figura 7 *Algoritmo de maximización.*

$$\text{Peso}(s_i) = \sum_{(r_j, w_j) \in R(w)} (r_j, w_j) \times \frac{\text{wnsf}(s_i, r_j)}{\sum_{s_i \in S(w)} \text{wnsf}(s_i, r_j)} \qquad (7)$$

$$\text{wnsf}(s_i, r_j) = \max_{s_x \in S(r_j)} (\text{pswn}(s_i, s_x))$$

wnsf (WordNet *similarity function*) es una medida de similitud semántica basada en WordNet que compara todos los sentidos de r_j con s_i obteniendo el sentido de r_j que mayor similitud tenga con s_i. Se ha utilizado la propuesta por Pedersen *et al*., *Extended Gloss Overlap*, la cual es una adaptación del algoritmo original de Lesk, medida que fue elegida por las siguientes razones:

- Obtuvo el mejor resultado en los experimentos realizados por Pedersen *et al*., quienes evaluaron diferentes medidas semánticas utilizando el paquete WordNet::Similarity.
- Pese a que en nuestros experimentos sólo se desambiguan sustantivos de

SENSEVAL-2, podríamos haber elegido una medida como la de Jiang-Conrath que trabaja sólo con las jerarquías de hipónimos y hiperónimos de sustantivos, cosa que no sucede con adjetivos y adverbios. En dicho caso sólo se podría trabajar con las palabras de la lista cuya categoría gramatical fuera sustantivo. En cambio, *Extended Gloss Overlap* trabaja sobre múltiples jerarquías sin tomar en cuenta la categoría gramatical de los vocablos.

Capítulo 5

Análisis experimental

SENSEVAL, es una organización dedicada a la investigación sobre el área de ambigüedad de sentidos de palabras. Su propósito es evaluar las debilidades y fortalezas de diferentes métodos que intentan resolver este fenómeno del lenguaje. SENSEVAL-2 es el segundo evento internacional que evalúa sistemas de desambiguación de palabras, el cual se llevó a cabo en Toulouse, Francia.

Los resultados de esta tesis se han comparado con los obtenidos en SENSEVAL-2, específicamente en la tarea *English all-words*, cuyos resultados se pueden ver en la tabla 4. Esta tarea consiste en asignar un sentido a cada una de los 2,473 vocablos ambiguos de un total de 5,000 palabras extraídas de artículos de Penn TreeBank y Wall Street Journal.

Orden	Sistema	Tipo	Precision	Recall	Cobertura (%)
1	SMUaw	supervised	0.690	0.690	100.00
	Mc.Carthy *et al.*	unsupervised	0.640	0.630	100.00
2	CNTS-Antwerp	supervised	0.636	0.636	100.00
3	Sinequa-LIA - HMM	supervised	0.618	0.618	100.00
–	*WordNet most frequent sense*	*supervised*	0.605	0.605	100.00
4	UNED - AW-U2	unsupervised	0.575	0.569	98.90
5	UNED - AW-U	unsupervised	0.556	0.550	98.90
6	UCLA - gchao2	supervised	0.475	0.454	95.55
7	UCLA - gchao3	supervised	0.474	0.453	95.55
9	CL Research - DIMAP (R)	unsupervised	0.451	0.451	100.00
10	UCLA - gchao	supervised	0.500	0.449	89.72

Tabla 4 *Resultados de SENSEVAL-2.*

La tercera columna de la tabla 4 muestra si un sistema es supervisado o no. Como se puede observar claramente, los mejores resultados corresponden a sistemas que aprenden de corpus de sentidos etiquetados manualmente (la mayoría toma en cuenta WordNet para este tipo de tareas), aunque esto puede resultar muy caro, sobre todo cuando no existen este tipo de recursos para un lenguaje específico. Por

lo general, sólo el idioma inglés presenta recursos etiquetados semánticamente de una manera confiable, tal como es el caso de SemCor. Nuestro método es no supervisado, por ende lo comparamos con métodos de las mismas características.

En la primera parte de esta sección se describen los recursos léxicos que se han utilizado en los experimentos (ver sección 5.1). Luego, se presenta de una manera muy general los tipos de experimentos planteados (ver sección 5.2). En las secciones 5.3 y 5.4 y 5.5 se detalla específicamente los tres tipos de experimentos realizados.

5.1 Recursos léxicos

En esta sección se describen los recursos léxicos u orígenes de información que se han utilizado en los experimentos realizados. Dependiendo del rol que cumple cada uno en el proceso de desambiguación, estos pueden ser divididos en tres categorías.

- Recursos planos. Para la creación de modelos de espacios de palabras usamos diferentes corpus textuales, tales como British National Corpus (100 millones de palabras), SemCor Corpus (1 millón de palabras) y Tasa Corpus (10 millones de palabras). Estos recursos tienen una característica común: presentan texto plano y dependiendo de cómo sean procesados, son ideales para la extracción de vocablos relacionados sintagmática y paradigmáticamente (ver figura 8).

- Recursos léxicos previamente compilados que contienen vocablos relacionados paradigmáticamente. Un ser humano puede deducir claramente vocablos de este tipo; por ende un tesauro manual es un buen repositorio para explorar. Se eligieron dos: el tesauro manual de Mobby y el tesauro de Lin, que a diferencia del anterior se construye automáticamente.

- Recursos léxicos previamente compilados que contienen vocablos relacionados sintagmáticamente. Como se explicó en secciones anteriores, el ejemplo más claro de esta relación son las colocaciones gramaticales. Se ha elegido el recurso más amplio que existe actualmente: El corpus de Google, el cual proporciona *n-gramas* obtenidos de Internet. La figura 9 nos muestra el rol de ambos tipos de recursos léxicos: Los sintagmáticos y los paradigmáticos.

–

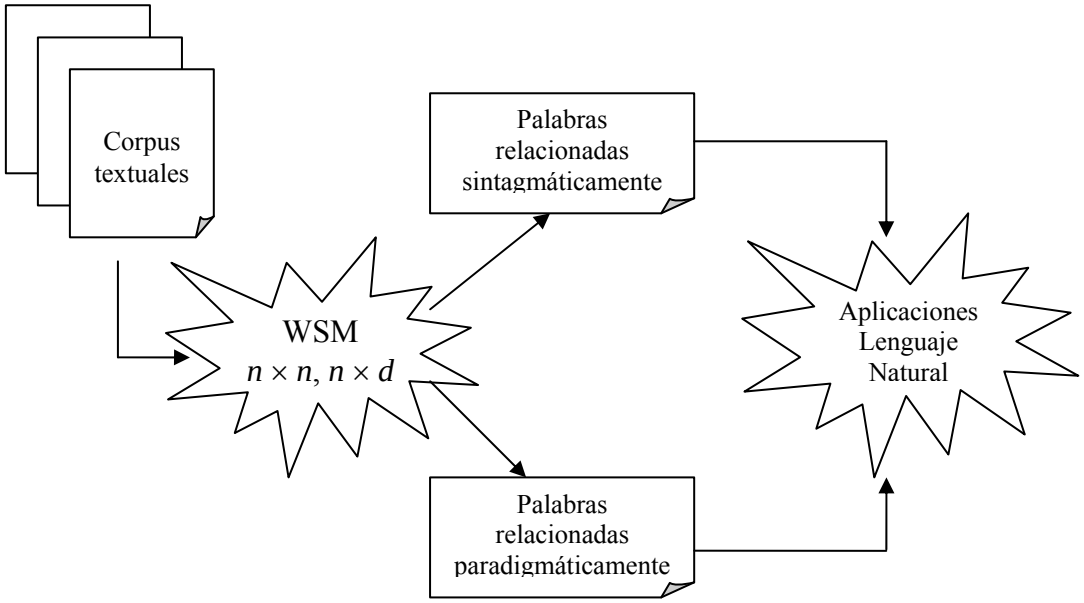


Figura 8 *Uso de corpus textuales como origen de información.*

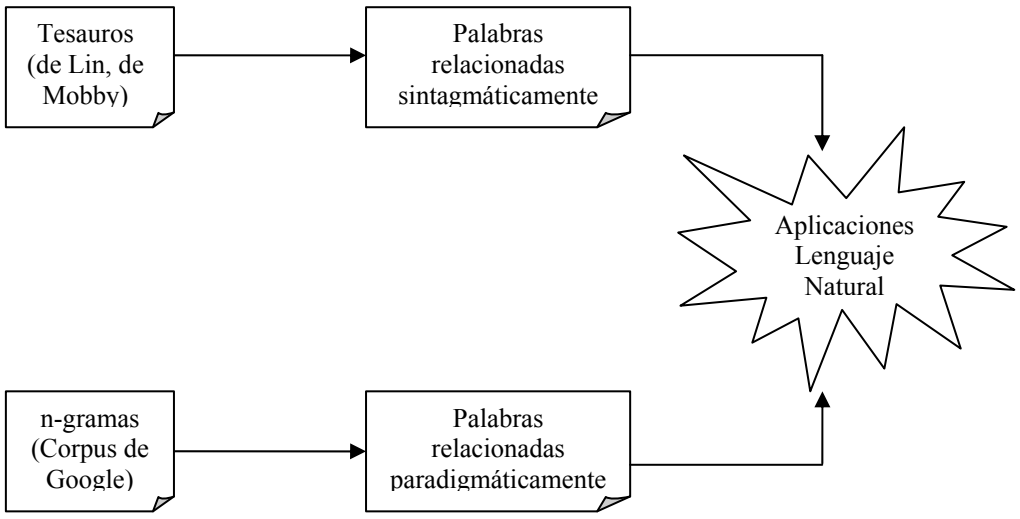


Figura 9 *Uso de recursos lingüísticos como origen de información.*

5.1.1 Tesauros

Los recursos léxicos que por excelencia proporcionan vocablos relacionados paradigmáticamente son los tesauros. Un tesauro es un conjunto de palabras con diferentes tipos de relaciones, tales como sinonimia, antonimia, homonimia,

hiperonimia, meronimia, etc. Existen tesauros que están orientados a temáticas específicas y otros abarcan un panorama más genérico. En el mundo de habla inglesa, es clásico el tesauro de Roget, cuya función es, según su autor, además de ayudar al escritor a encontrar la palabra que exprese mejor su pensamiento, también estimular su intelecto y sugerirle palabras o ideas relacionadas.

Algunos autores argumentan que un tesauro es el producto final de un modelo de espacio de palabras. En [54], se demostró que la intersección entre la información paradigmática y sintagmática que proporcionan diversos WSM y la de un tesauro que se crea manualmente es muy baja: 10%. En dicho trabajo se concluye que un WSM proporciona conjuntos de palabras relacionadas semánticamente; sin embargo aún no se puede determinar específicamente el tipo o clase de esa relación

Los términos que conforman un tesauro se interrelacionan entre ellos bajo tres modalidades de relación:

- Relaciones jerárquicas. Establece subdivisiones que generalmente reflejan estructuras de todo/parte.
- Relaciones de equivalencia. Controla la sinonimia, homonimia, antonimia y polinimia entre los términos.
- Relaciones asociativas. Mejoran las estrategias de recuperación y ayudan a reducir la poli jerarquía entre los términos.

En este trabajo se han elegido dos tesauros cuya diferencia radica en la manera como han sido construidos. El tesauro de Mobby se construyó manualmente y el de Lin automáticamente [34]. A continuación se describe cada uno de ellos.

a. Tesauro de Mobby

El tesauro de Mobby está considerado como una de las fuentes de información más grandes y coherentes que existen para el idioma inglés. Este tesauro es manual, es decir, creado por el ser humano. La segunda edición de dicho tesauro ha mejorado mucho con respecto a la primera, añadiendo mas de 5000 palabras principales, que en su totalidad superan los 30000 vocablos.

Asimismo, se le añadió también más de un millón de sinónimos y *términos relacionados*, que en su totalidad superan los 2.5 millones de sinónimos y *términos relacionados*. La figura 10 muestra una entrada del tesauro.

demon: baba yaga, lilith, mafioso, satan, young turk, addict, afreet, ape-man, atua, barghest, beast, beldam, berserk, berserker, bomber, brute,bug, cacodemon, collector, daemon, daeva, damned spirits, demonkind, demons, denizens of hell, devil, devil incarnate, dragon, dybbuk, eager beaver, energumen, enthusiast, evil genius, evil spirit, evil spirits, faddist, fanatic, fiend, fiend from hell, fire-eater, firebrand, freak, fury, genie, genius, ghoul, goon, gorilla, great one for, gunsel, gyre, hardnose, harpy, hell-raiser, hellcat, hellhound, hellion, hellish host, hellkite, hobbyist, holy terror, hood, hoodlum, host of hell, hothead, hotspur, hound, incendiary, incubus, infatuate, inhabitants of pandemonium, intelligence, jinni, jinniyeh, killer, lamia, lost souls, mad dog, madcap, monster, mugger, nut, ogre, ogress, powers of darkness, pursuer, rakshasa, rapist, revolutionary, rhapsodist, satan, savage, she-wolf, shedu, souls in hell, specter, spirit, spit_re, succubus, sucker for, supernatural being, termagant, terror, terrorist, the damned, the lost, the undead, tiger, tigress, tough, tough guy, ugly customer, vampire, violent, virago, visionary, vixen, werewolf, wild beast, witch, wolf, yogini, zealot

Figura 10 *Vocablo demon en el tesauro de Mobby*

b. Tesauro de Lin

Este tesauro, creado automáticamente, fue uno de los primeros recursos léxicos que se construyó usando una técnica de lenguaje natural, conocida como *Bootstrapping semantics* [34]. Para ello, primero se define una medida de similitud que se basa en los patrones de distribución de las palabras. Esta medida permite construir un tesauro usando un corpus parseado, el cual consta de 64 millones de palabras, las cuales fueron tomadas de diversos corpus tales como el Wall Street Journal (24 millones de palabras), San Jose Mercury (21 millones de palabras) y AP Newswire (19 millones de palabras). Una vez que el corpus fue parseado se extrajo 56.5 millones de tripletas de dependencia (de las cuales 8.7 millones fueron únicas). En el corpus parseado hubo 5469 sustantivos, 2173 verbos, y 2632 adjetivos/adverbios que ocurrieron al menos 100 veces.

Finalmente, se computó la similitud semántica de cada par de vocablos entre sustantivos, todos los verbos, adjetivos y adverbios. Para cada una de estas palabras se

creó una entrada en el tesauro, que contiene una *lista ponderada* de términos similares.

5.1.2 Corpus de Texto

Los corpus que se describen en esta sección, son los que se han usado en la construcción de diferentes modelos de espacios de palabras. El primero de ellos, British National Corpus (100 millones de palabras) es 10 veces más grande que Tasa Corpus (10 millones). Actualmente, en lo referente a la construcción de WSM, se prefiere los recursos más grandes ya que estos proporcionan mejor fundamento estadístico; sin embargo en [52] se analizó la densidad de los WSM creados con ambos tipos de corpus, y se concluyó que los corpus muy extensos como BNC tienden a distorsionar las relaciones semánticas entre los vocablos, mientras que WSM creados con corpus no tan grandes, presentan un mejor rendimiento en lo referente a dicha característica.

a. British National Corpus (BNC)

British National Corpus (BNC) es el más extenso de los corpus que existen actualmente. Está conformado por una colección de 100 millones de palabras de ejemplos hablados y escritos tomados de una amplia gama de recursos léxicos. BNC está diseñado para representar el inglés británico de la última década del siglo XX (tanto el escrito como el hablado). La última edición de este recurso es *BNC XML Edition* y fue liberada en el año 2007.

La parte escrita de BNC (90%), por ejemplo incluye, extracto de periódicos nacionales y regionales, revistas especializadas y folletos para todas las edades e intereses, ensayos escolares y universitarios, entre otras muchas clases de texto. La parte hablada (10%) se encuentra conformada por transcripciones ortográficas de conversaciones informales (grabadas por voluntarios de diferentes edades, regiones y clases sociales distribuidos uniformemente en una región). Asimismo, contiene lenguaje hablado, el cual se utiliza en diferentes contextos, desde negocios formales, reuniones de gobierno hasta programas radiales y llamadas telefónicas.

b. Corpus Tasa

El corpus Tasa (Touchstone Applied Science Associates) esta conformado por textos de nivel secundario los cuales comprenden diferentes tópicos, tales como lenguaje, artes, salud, economía, ciencia, estudios sociales y negocios. Usamos este corpus por dos razones:

- Esta dividido en secciones de aproximadamente 150 palabras. Esto significa que es un recurso *ad-hoc* para usarlo como origen de información sintagmática; ya que cada una de estas secciones representaría a un documento en las columnas de la matriz. Otros corpus como BNC no se encuentra divido en regiones lógicas como Tasa Corpus.

- Otra de las razones es su tamaño. Es lo suficientemente grande como para obtener estadísticas confiables y no es lo suficientemente pequeño como para crear modelos de espacios de palabras sin semántica.

5.1.3 Otros corpus

En esta categoría se describe dos corpus que no tienen semejanza con los descritos anteriormente. SemCor es un corpus pequeño (menor a 1 millón de palabras), donde cada uno de sus vocablos ha sido etiquetado semánticamente con los sentidos proporcionados por WordNet. Generalmente se usa para evaluar métodos orientados a solucionar la desambiguación de sentidos de palabras. El otro corpus que usamos es proporcionado por Google y es uno de los más extensos en lo que se refiere a *n-gramas* (vocablos relacionados sintagmáticamente). Usamos el corpus de Google para extraer vocablos relacionados sintagmáticamente, al igual que un WSM sintagmático, con la finalidad de evaluar su calidad semántica en el método de desambiguación de sentidos de palabra que se detalla en el siguiente capítulo.

a. Corpus Semcor

SemCor es un corpus léxico etiquetado semánticamente, el cual fue creado por la universidad de Princeton. Éste es un subconjunto del corpus *English Brown*. Actualmente, SemCor contiene al menos 700,000 palabras etiquetadas con su categoría

gramatical y más de 200,000 palabras son proporcionadas con su respectivo lema y número de sentido tomando como referencia WordNet. Las palabras cuya categoría gramatical hace referencia a preposiciones, determinantes, pronombres y verbos auxiliares no son etiquetadas semánticamente, al igual que caracteres no alfanuméricos, interjecciones y términos coloquiales.

Más en detalle SemCor consta de un total de 352 archivos. En 186 de ellos, sustantivos, adjetivos, verbos y adverbios son etiquetados con su categoría gramatical, lema y sentido; mientras que en los 166 textos restantes, sólo los verbos son etiquetados con su lema y sentido. El número total de *tokens* en SemCor es de 359,732 en el primer conjunto de archivos, de los cuales 192,639 han sido etiquetados semánticamente, mientras que el segundo grupo, este número asciende a 316,814 *tokens*, de los cuales 41,497 ocurrencias de verbos han sido etiquetadas semánticamente.

b. Corpus de Google

El corpus de Google, el cual es distribuido por Linguistic Data Consortium (www.ldc.upenn.edu), es quizás el recurso más extenso de *n-gramas* que existe en la actualidad. Dicho corpus consta de archivos de texto (200 a 300 megabytes cada uno) dispersos en 6 DVDs. Estos almacenan información referente a bigramas, trigramas, tetragramas y pentagramas. Esta información ha sido recopilada por Google directamente de Internet, por ende la amplitud de temas que reflejan sus *n-gramas* es muy amplia en cantidad y contenido. Debido a lo extenso del recurso y a la forma como esta organizado (archivos planos) las consultas o búsqueda de información se hace lenta y difícil. Por ende se ha creado una base de datos SQL Server que almacena dicha información con la finalidad de optimizar este proceso. El acceso a dicha base de datos está disponible en 148.204.20.174/Google_corpus/default.aspx.

5.2 Tipos de experimentos

Los experimentos realizados se clasifican en tres tipos, los que detallamos en esta sección, desde un punto de vista muy general.

– En el primero, se construyen WSM que proporcionan *términos relacionados*

sintagmática y paradigmáticamente. Algunos parámetros, tales como el tamaño de la ventana contextual (en el caso de los paradigmáticos) y la región de contexto (en el caso de los sintagmáticos) empleados en la construcción han sido analizados. Asimismo, el tipo de contexto de la instancia ambigua utilizado en la selección de *términos relacionados* también ha sido explorado. El objetivo es determinar cuál de los dos tipos de WSM maximiza los resultados del método de desambiguación planteado.

– En el segundo, se reemplaza el WSM por otros recursos léxicos, como el tesauro de Lin, el tesauro de Mobby y el corpus de Google, describiendo la metodología de selección de *términos relacionados* que se ha implementado para cada uno. El objetivo de este experimento es determinar cuál de estos recursos es el que maximiza los resultados del método planteado. Asimismo, se comparan estos resultados con los obtenidos en el primer experimento. De esta manera, se determina si es necesario crear un WSM para obtener un conjunto de *términos relacionados*, o sólo basta con utilizar alguno de los tesauros o corpus mencionados.

– En el tercero, se analiza el impacto del método planteado no sólo en la detección del sentido de un vocablo en un contexto específico, sino también en la detección del sentido predominante. En la misma sección, se analiza también el impacto de la temática del corpus al proporcionar *términos relacionados* que se aplican a desambiguar vocablos del mismo corpus o a otros de diferentes corpus.

5.3 WSM sintagmáticos y paradigmáticos

El método de desambiguación propuesto puede ser aplicado a cualquier lenguaje. Se ha elegido el idioma inglés, ya que las herramientas de análisis morfológico y sintáctico, corpus, tesauros, diccionarios computacionales, y demás recursos externos tienen mayor confiabilidad que los existentes en otros idiomas. En este trabajo sólo se han desambiguado sustantivos, dejando de lado los adjetivos y adverbios. En el caso de los verbos, los resultados no son prometedores, ya que tienen muchos sentidos y las diferencias entre sus glosas (definición de un sentido en WordNet) son muy sutiles.

Como se detalló en esta sección, la primera etapa del experimento concierne a la construcción de WSM. Se han construido tres WSM: Un sintagmático y un paradigmático tomando la información de BNC como origen de información, y otro paradigmático tomando la información de Semcor como fuente de entrenamiento. En los WSM paradigmáticos se ha tomado una *ventana contextual* de tres vocablos a la izquierda y tres a la derecha y en los sintagmáticos el tamaño de la *región contextual* es de 20 vocablos.

En el caso del corpus Semcor, no se decidió construir un WSM sintagmático debido al exiguo tamaño de dicho corpus (menos de 800,000 vocablos). Asimismo, en la construcción de su WSM paradigmático sólo se usó el 90% del texto en la construcción y se reservó el 10% para su evaluación. Esto se realizó con la finalidad de analizar el impacto de un WSM en una temática específica (esto se detalla en la sección 5.5.2)

Una vez creados dichos WSM, la segunda etapa consiste en extraer el conjunto de *términos relacionados* con la instancia ambigua y aplicar el algoritmo de maximización. Para ello, se crea un vector con el contexto local del vocablo ambiguo, el cual se compara con cada uno de los vectores almacenados en el modelo de espacio de palabras.

El rol del contexto es una variable que influye directamente en la selección de *términos relacionados*. En los resultados que se presentan a continuación se han usado dos tipos de contexto claramente diferenciados: Las dependencias sintácticas (ver tabla 5 y gráfica 1) y las ventanas de contexto de amplitudes variables a la derecha y a la izquierda de la instacia ambigua (ver tabla 6 y gráficas 2 y 3), tales como dos vocablos a la izquierda y dos a la derecha (2-2), tres a la izquierda y tres a la derecha (3-3), cero a la izquierda y tres a la derecha (0-3).

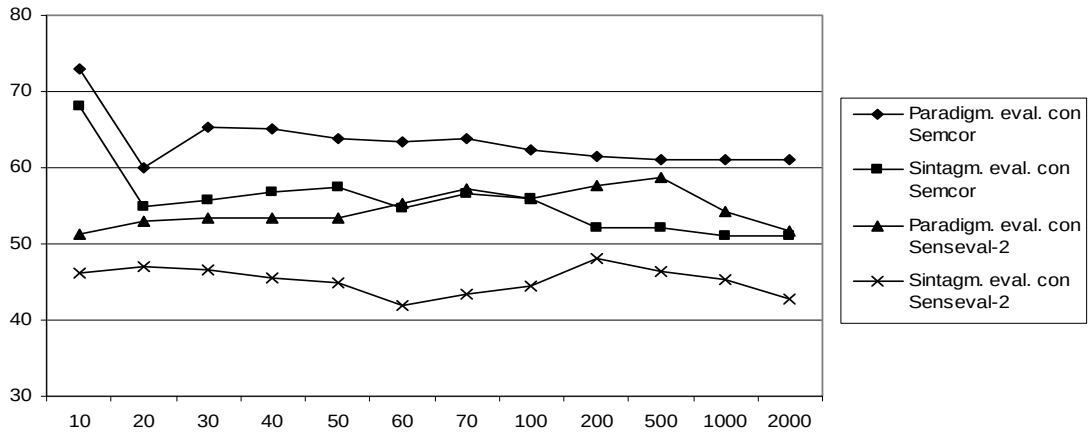
La gráfica 1 muestra los resultados obtenidos por el algoritmo de maximización cuando los *términos relacionados* son proporcionados por el WSM sintagmático y paradigmático, construidos con la información de BNC. La evaluación se realiza con la totalidad de Semcor y SENSEVAL-2.

Como se aprecia en la tabla 5, cuando se entrena con BNC y se evalúa con el

corpus *English all-words* SENSEVAL-2 el promedio correspondiente a la *precision* obtenida cuando se procesan vocablos paradigmáticos es de 54.60% contra 45.20% de los sintagmáticos y cuando se evalúa con Semcor los resultados son de 58.38% contra 55.53% respectivamente. Estos resultados son un claro indicador que los vocablos proporcionados por los Modelos Paradigmáticos tienen mayor impacto en el método de desambiguación planteado. Asimismo, en la gráfica 1, también se puede apreciar la supremacía de los modelos paradigmáticos sobre los sintagmáticos

Evaluado con:		SENSEVAL-2			Semcor		
Entrenado con:		SemCor	BNC		SemCor	BNC	
Tipo WSM		Paradig.	Paradig.	Sintag.	Paradig.	Paradig.	Sintag.
n-ésimos *términos relacionados*	**10**	44.22	51.35	46.09	64.23	**73.07**	**68.09**
	20	44.77	52.88	47.03	69.44	60.00	54.86
	30	45.91	53.33	46.54	67.36	65.27	55.76
	40	45.76	53.33	45.54	66.43	65.16	56.76
	50	45.55	53.33	44.88	67.80	63.80	57.53
	60	48.12	55.36	41.98	68.15	63.41	54.65
	70	**49.84**	57.22	43.34	**69.86**	63.84	56.54
	100	48.80	56.02	44.55	**69.86**	62.33	55.90
	200	49.05	57.57	**48.00**	66.75	61.58	52.10
	500	49.10	**58.79**	46.29	65.89	61.08	52.10
	1000	44.55	54.27	45.29	65.06	61.08	51.05
	2000	41.05	51.75	42.87	62.76	61.08	51.01
Promedio		42.97	54.60	45.20	61.73	58.38	55.53

Tabla 5 *Usos de dependencias sintácticas como contexto local.*



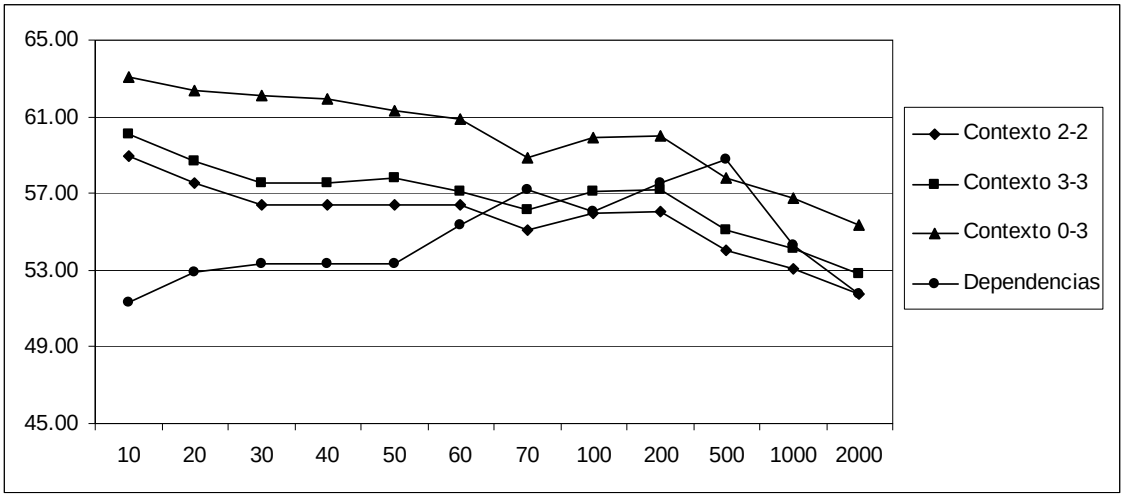
Gráfica 1 *Resultados de WSD, usando WSM entrenado con BNC, y dependencias sintácticas como contexto de la instancia ambigua.*

En la tabla 6 y la gráfica 2 y la gráfica 3, muestran el impacto que puede tener el contexto del vocablo ambiguo en la selección de *términos relacionados*. En este experimento se usó sólo BNC para la construcción de un WSM sintagmático y otro paradigmático. Asimismo, se usó diferentes tipos de contexto de cada una de las instancias ambiguas del corpus de evaluación *English all-words* SENSEVAL-2.

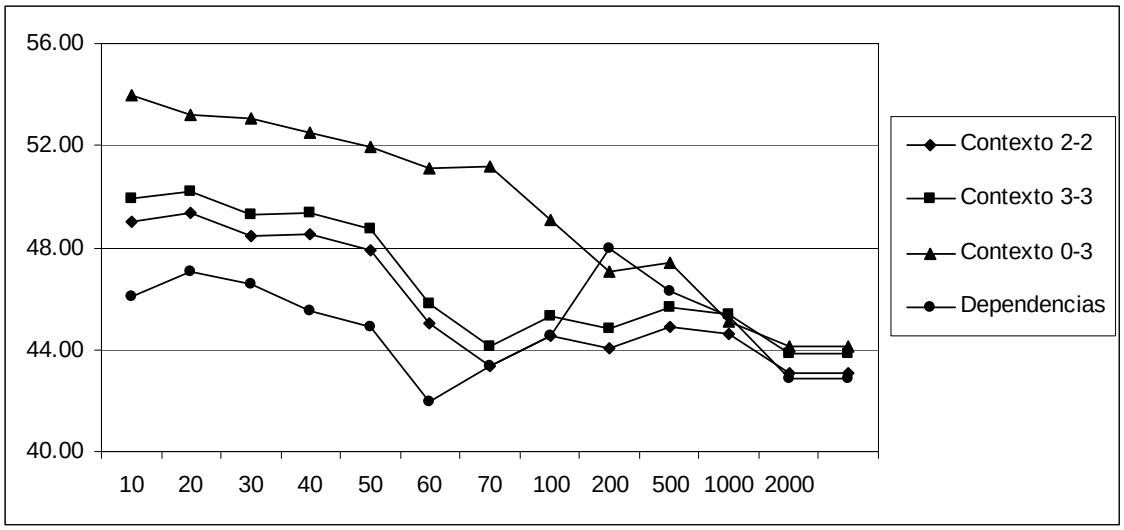
Contexto		2-2		3-3		0-3		Dependencias	
Tipo WSM		Paradig.	Sintag.	Paradig.	Sintag.	Paradig.	Sintag.	Paradig.	Sintag.
n-ésimos términos relacionados	10	58.92	49.04	60.10	49.90	63.04	53.98	51.35	46.09
	20	57.55	49.33	58.70	50.20	62.34	53.23	52.88	47.03
	30	56.43	48.44	57.56	49.29	62.07	53.08	53.33	46.54
	40	56.40	48.54	57.53	49.39	61.89	52.54	53.33	45.54
	50	56.40	47.88	57.81	48.72	61.34	51.98	53.33	44.88
	60	56.40	45.03	57.09	45.82	60.88	51.09	55.36	41.98
	70	55.05	43.33	56.15	44.09	58.90	51.15	57.22	43.34
	100	55.98	44.55	57.10	45.33	59.90	49.08	56.02	44.55
	200	56.04	44.07	57.16	44.85	59.96	47.09	57.57	48.00
	500	54.01	44.90	55.09	45.69	57.79	47.40	58.79	46.29
	1000	53.06	44.60	54.12	45.38	56.77	45.10	54.27	45.29
	2000	51.75	43.09	52.79	43.85	55.37	44.09	51.75	42.87
Promedio		55.67	46.07	56.77	46.88	**60.02**	**49.98**	54.60	45.20

Tabla 6 *Uso de ventanas variables como contexto local.*

En la tabla 6, se puede apreciar que los mejores resultados se obtienen usando un contexto de cero vocablos a la izquierda y tres a la derecha. El segundo y tercer lugar corresponden a las ventanas de contexto simétricas 2-2 y 3-3. Finalmente, el contexto de menor rendimiento es el correspondiente a las dependencias sintácticas.



Gráfica 2 *Resultados de WSD, usando WSM Paradigmático entrenado con BNC, y evaluado con SENSEVAL-2.*



Gráfica 3 *Resultados de WSD, usando WSM Sintagmático entrenado con BNC, y evaluado con SENSEVAL-2.*

En las gráficas 2 y 3, además de confirmar a la ventana contextual de 0-3 como la de mejor rendimiento, también se observa que es mejor utilizar un número reducido de *términos relacionados*. Asimismo, dichas gráficas muestran la clara ventaja de los WSM paradigmáticos sobre los sintagmáticos cuando los *términos* que proveen se utilizan en el método de WSD planteado. Una característica que es importante resaltar, es la naturaleza del corpus de entrenamiento. Claramente se puede apreciar, que los resultados bajan drásticamente cuando se evalúa con el corpus *English all-words* SENSEVAL-2, independientemente del corpus de entrenamiento que se utilice. Esto nos indica que los sustantivos de dicho corpus son más *difíciles de desambiguar* que los existentes en Semcor.

En esta sección se presenta aquellos sustantivos que siempre se desambiguaron mal o bien, tal como se muestra en la tabla 7; sin embargo el promedio correspondiente al número de sentidos de dichos sustantivos se ha calculado sobre la totalidad. El promedio del número de sentidos de los sustantivos que siempre se desambiguaron mal es mayor. (6.21 vs. 3.33)

Sobre el número de elementos de la *lista ponderada*, el mejor resultado, 73.07%, se obtuvo cuando se usaron los primeros 10 elementos de la *lista ponderada*. Asimismo,

cuando el número de elementos se incrementa, los resultados decrementan, de lo cual se concluye que el algoritmo de maximización propuesto por Mc. Carthy obtiene mejores resultados cuando se usa un pequeño conjunto de *términos relacionados*. Los resultados obtenidos por Mc. Carthy (48% en precisión) son muy inferiores a los que hemos obtenido. Esto se debe a que dicho método, a diferencia del nuestro, no toma en cuenta el contexto de los vocablos ambiguos en la selección de los elementos de una *lista ponderada*.

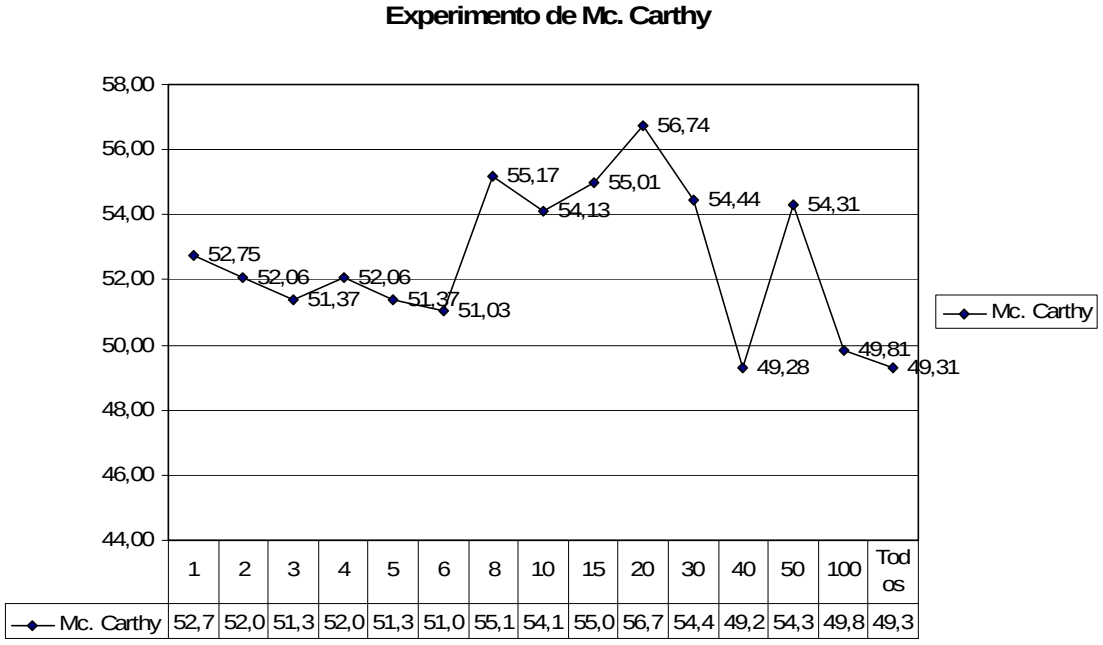
Palabras que siempre se desambiguaron bien		Palabras que siempre se desambiguaron mal	
Palabra	Nro. de Sentidos	Palabra	Nro. de Sentidos
woman	4	ringer	4
growth	7	loss	8
teacher	2	performance	5
Eye	5	sort	4
Rope	2	time	10
Team	2	form	16
Polyp	2	ringers	4
individual	2	life	14
Leader	2	role	4
women	4	politician	3
Skin	7	influence	5
City	3	change	10
treatment	4	scale	10
frequency	3	one	2
bronze	2	pap	3
experiment	3	swing	9
mother	5	note	9
doctor	4	thing	12
copies	4	source	9
personality	2	passion	7
structure	5	parliament	2
public	2	authority	7
purpose	3	posture	4
ropes	2	fault	7
nation	4	array	4
Right	8	tests	6
polyps	2	situation	5
evening	3	exposure	10
childhood	2	identity	4
teams	2	print	6
Promedio	3.33		6.21

Tabla 7 *Vocablos que siempre se desambiguaron bien o mal.*

El peor resultado, 41.05%, se produjo cuando se utilizaron 2000 elementos en la *lista ponderada*. Este resultado permite afirmar que mientras el conjunto de vocablos

relacionados es más numeroso, los resultados obtenidos por el algoritmo empeoran. Para determinar el comportamiento del algoritmo de desambiguación propuesto por Mc.Carhty *et al*., se reprodujo dicho experimento y se analizó su comportamiento (ver gráfica 4).

Mc. Carhty *et al*., a diferencia nuestra, no construyó ningún modelo de espacio de palabras. En su algoritmo de desambiguación, para cada instancia ambigua consulta al tesauro de Lin, y toma el conjunto de elementos relacionados que éste le proporciona. Finalmente, aplica su algoritmo de maximización. Al experimentar sobre los vocablos ambiguos de SENSEVAL-2 reportó 48% de precisión. Nosotros, al reproducir sus experimentos sólo desambiguamos los sustantivos de SENSEVAL-2, obteniendo una precisión de 49.3%. La pequeña diferencia (1.3%) que existe con sus resultados se debe a que ella experimentó con todas la palabras polisémicas de SENSEVAL-2, incluídos los verbos, que son mucho más polisémicos que los sustantivos; por lo cual sus resultados son un poco más bajos.



Gráfica 4 *Comportamiento del algoritmo de Mc. Carthy et al.*

En la gráfica 4, claramente se puede observar que los resultados de Mc. Carthy *et al*. se vieron perjudicados por incluir la totalidad de las palabras proporcionadas por el tesauro de Lin. El mejor resultado, 56.74%, se obtiene cuando sólo se incluyen 20

términos relacionados, y el segundo, 55.17%, y tercer mejor resultado, 55.01%, se obtienen cuando sólo se incluyen con 8 y 15 *términos relacionados* respectivamente. Estos resultados nos permiten afirmar que un vocablo ambiguo sólo necesita un pequeño grupo de palabras fuertemente relacionadas para que su sentido pueda ser determinado por el algoritmo de maximización.

5.4 WSM y otros recursos léxicos

En los experimentos anteriores se construyeron modelos de espacio de palabras, a los que se consulta con la finalidad de obtener una lista de *términos relacionados* sintagmática y paradigmáticamente con una instancia ambigua. Los elementos de dicha *lista* se seleccionan tomando en cuenta el contexto local del vocablo ambiguo.

La calidad semántica de los *términos* depende en gran parte del corpus que se usa en la construcción del WSM y de las *ventanas contextuales* (WSM paradigmático) y *regiones de contexto* (WSM sintagmático) empleados.

Actualmente, aún no se ha determinado clara y específicamente el tipo de información que proporciona un WSM; sin embargo podemos comparar dicha información con la que proporcionan otros recursos más específicos y aplicarlas en tareas del procesamiento de lenguaje natural, más específicamente a WSD. De esta manera podemos determinar la utilidad de la información que proporciona un WSM sintagmático o paradigmático con respecto a la suministrada por otros recursos, específicamente: El tesauro de Mobby y el tesauro de Lin. El objetivo principal de los experimentos que se presentan en esta sección es determinar cuál de estos recursos proporcionan los mejores *términos relacionados*, los cuales son fundamentales en el método de desambiguación planteado.

En [55], se comparó la información paradigmática y sintagmática proporcionada por diferentes WSM con la existente en el tesauro manual de Mobby. Los resultados muestran que la información paradigmática contenida en el WSM, con respecto al tesauro de Mobby, es de 7.78%, mientras que la sintagmática es de 10.43%. En dichos experimentos, se llegó a la conclusión que los WSM sintagmáticos contienen vocablos relacionados sintagmática y paradigmáticamente mientras que los WSM paradigmáticos

contiene información más homogénea, es decir, términos explícitamente relacionados paradigmáticamente.

5.4.1 Tesauro de Mobby

Al consultar un WSM y obtener una *lista ponderada* para una instancia ambigua, se toma en cuenta el contexto del vocablo ambiguo. Para ello, se crea un vector de dicho contexto y se le compara con todos los vectores del WSM, seleccionado los vectores más cercanos y por ende las palabras más similares a la instancia ambigua. Una *entrada* o vocablo en el tesauro de Mobby es un conjunto de palabras listadas alfabéticamente, que carecen de algún valor o peso de ponderación que determine el grado de similitud o relación semántica con dicha *entrada*. (ver sección 3.1.1). Para ello, se ha creado un método que tome en cuenta el contexto en la selección de *términos relacionados* proporcionados por dicho tesauro, el cual se detalla a continuación:

1. Obtener el contexto del vocablo ambiguo, el cual está conformado por un conjunto de vocablos.

$$C(w) = \{c_1, c_2, \ldots, c_i\}, \tag{8}$$

   Donde *w* es la instancia ambigua y c_i es cada uno de los vocablos que pertenecen a su contexto.

2. Obtener el conjunto de vocablos relacionados semánticamente con la palabra ambigua, los cuales son proporcionados por el tesauro de Mobby.

$$S(w) = \{sw_1, sw_2, \ldots, sw_i\}, \tag{9}$$

   Donde *w* es el vocablo ambiguo y sw_i es cada una de las palabras que proporciona el tesauro.

3. Discriminar los vocablos obtenidos en S(w) tomando en cuenta el contexto de la instancia ambigua. Para ello, se construye un WSM (ver tabla 8) donde cada columna es una palabra de C(w) y cada fila, un miembro de S(w). Asimismo, se incluye el vocablo ambiguo *w* como una fila más.

	c_1	c_2	c_3	c_i
sw_1	(c_1, sw_1)	(c_2, sw_1)	(c_3, sw_1)	(c_i, sw_1)
sw_2	(c_1, sw_2)	(c_2, sw_2)	(c_3, sw_2)	(c_i, sw_2)
sw_3	(c_1, sw_3)	(c_2, sw_3)	(c_3, sw_3)	(c_i, sw_3)
sw_j	(c_1, sw_j)	(c_2, sw_j)	(c_3, sw_j)	(c_i, sw_j)
w	(c_1, w)	(c_2, w)	(c_3, w)	(c_i, w)

Tabla 8 *WSM usando tesauro de Mobby.*

Cada par ordenado representa la máxima similitud semántica entre c_i y sw_i o en su caso, *w*. Este valor se computa comparando todos los sentidos de c_i con los de sw_i (ver figura 11) y se elige el más alto. La comparación de dos sentidos se realiza mediante el paquete WordNet::Similarity [46], específicamente con la medida *extended gloss overlap*, que no es otra cosa que una modificación del algoritmo de Lesk.[30].

4. Obtener los *términos relacionados* de *w*. En el WSM, cada vocablo sw_j se representa mediante un vector, al igual que *w*. De esta manera se compara el vector de *w* con todos los demás. La similitud de dos vectores en un espacio de *n* dimensiones se computa en base a sus dimensiones. Este cálculo se puede realizar de varias formas. En este trabajo se usó el valor del coseno del ángulo que forman ambos vectores (ver ecuación 5) como indicador de su proximidad o lejanía semántica. De esta manera, seleccionando los vectores más similares al de *w*, se obtiene los *términos relacionados*.

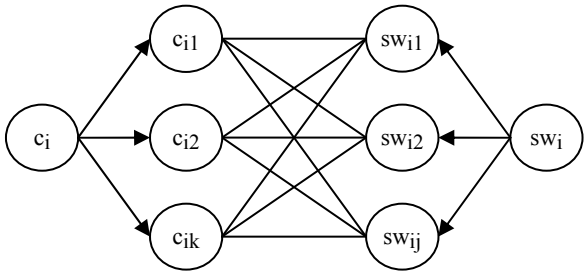


Figura 11 *Comparación de los sentidos de dos vocablos.*

5. Aplicar el algoritmo de maximización, utilizando el conjunto de *términos*

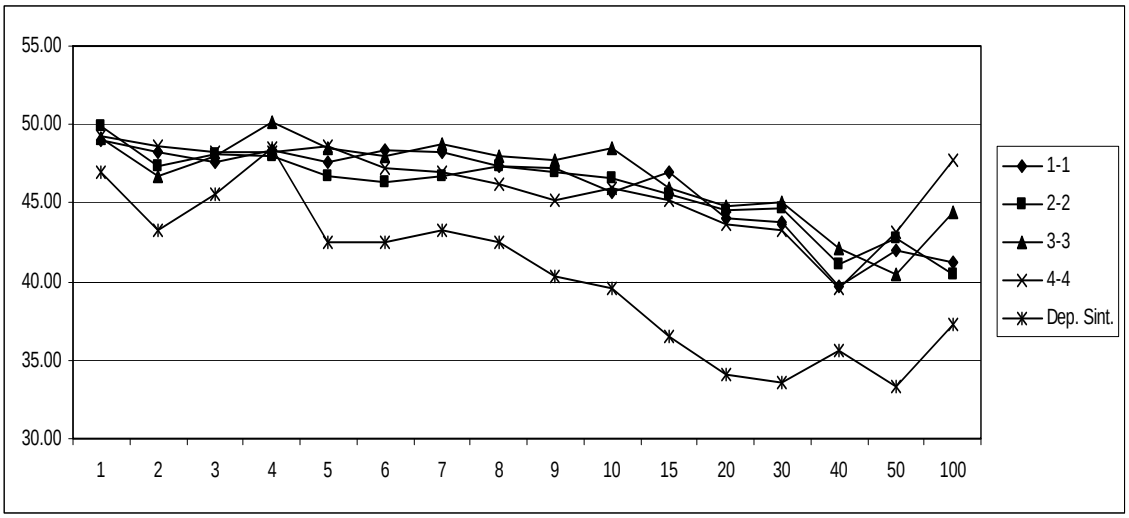
relacionados obtenidos en el punto anterior

En los experimentos que se detallan en esta sección, se han usado diferentes contextos de la instancia ambigua *w*, para la creación de su respectivo vector, ya que dicho contexto condiciona la selección de los *términos relacionados*. Los tipos de contexto utilizados han sido dependencias sintácticas y ventanas contextuales de diferente amplitud, tales como: 4 izq.–4 der., 3 izq.–3 der., 2 izq.–2 der., 1 izq.–1 der., y asimétricas: 3 izq.–0 der., 3 izq.–1 der., 3 izq.–2 der., 0 izq.–3 der., 1 izq.–3 der., 2 izq.–3 der.

Asimismo se han utilizando los primeros 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, y 100 *términos relacionados* con la instancia ambigua. La tabla 9 resume los experimentos que se realizaron, tomando en cuenta el tipo de contexto y número de *términos relacionados* usados. La gráfica 5 y la tabla 10 muestran los resultados cuando se usan ventanas simétricas y de dependencia sintácticas.

N° términos	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100
Simétricas																
1-1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2-2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3-3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4-4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Asimétricas																
0-3	✓	✓	✓	✓												
1-3	✓	✓	✓	✓												
2-3	✓	✓	✓	✓												
3-0	✓	✓	✓	✓												
3-1	✓	✓	✓	✓												
3-2	✓	✓	✓	✓												
Dependencias	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tabla 9 *Resumen de experimentos realizados.*



Gráfica 5 *Uso de ventanas simétricas y los primeros n-términos relacionados.*

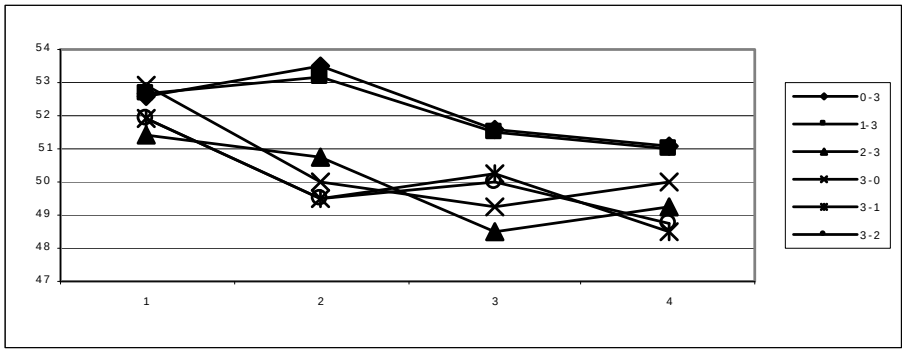
Nº términos	1-1		2-2		3-3		4-4		Dependencias Sintácticas		Promedio
	Evaluadas	Precisión	Evaluadas	Precisión	Evaluadas	Precisión	Evaluadas	Precisión	Evaluadas	Precisión	
1	622	49.04	507	**49.90**	379	49.08	292	**49.32**	134	47.01	**48.87**
2	622	48.23	507	47.34	379	46.7	292	48.63	134	43.28	46.84
3	622	47.59	507	48.13	379	48.02	292	48.29	134	45.52	47.51
4	622	48.39	507	47.93	379	**50.13**	292	48.29	134	48.51	**48.65**
5	622	47.59	507	46.75	379	48.55	292	48.63	134	42.54	46.81
6	622	48.39	507	46.35	379	48.02	292	47.26	134	42.54	46.51
7	622	48.23	507	46.75	379	48.81	292	46.92	134	43.28	46.80
8	507	47.34	507	47.34	379	48.02	292	46.23	134	42.54	46.29
9	622	47.27	507	46.94	379	47.76	292	45.21	134	40.3	45.50
10	622	45.66	507	46.55	379	48.55	292	45.89	134	39.55	45.24
15	622	46.95	507	45.56	379	45.91	292	45.21	134	36.57	44.04
20	620	44.03	505	44.55	377	44.83	291	43.64	132	34.09	42.23
30	616	43.83	502	44.62	375	45.07	289	43.25	131	33.59	42.07
40	580	39.66	472	41.10	349	42.12	270	39.63	129	35.66	39.63
50	555	41.98	454	42.73	336	40.48	260	43.08	123	**33.33**	40.32
100	417	41.25	341	40.47	257	44.36	201	47.76	94	37.23	42.21
Promedio		45.96		45.81		**46.65**		46.08		40.35	44.97

Tabla 10 *Resultados método de desambiguación tomando n vecinos del tesauro de Mobby y ventanas de contexto simétricas.*

La gráfica 6 y la tabla 11 muestran los resultados cuando se usan ventanas asimétricas tomando como límite tres vocablos a la izquierda y tres a la derecha.

Nº términos	0-3	1-3	2-3	3-0	3-1	3-2	Promedio
1	52.55	52.70	51.45	52.94	51.94	51.94	**52.25**
2	**53.52**	53.20	50.72	50.00	49.51	49.51	51.08
3	51.58	51.47	48.54	49.26	50.24	50.00	50.18
4	51.09	50.98	49.27	50.00	48.54	48.78	49.78
Promedio	52.19	52.09	50.00	50.55	50.06	50.06	50.83
Promedio por lado		**51.42**			50.22		

Tabla 11 *Método de desambiguación tomando ventanas de contexto asimétricas con un límite máximo de tres vocablos por lado.*



Gráfica 6 *Método de desambiguación tomando ventanas de contexto asimétricas con un límite máximo de 3 vocablos por lado.*

Los mejores resultados se obtienen cuando el algoritmo de maximización procesa menor cantidad de *términos relacionados* de la lista de ponderada que proporciona el tesauro de Mobby. Analizando el número de *términos relacionados* que se han utilizado en este experimento, se puede afirmar:

- En las ventanas simétricas, el mejor resultado fue 50.13% usando cuatro *términos relacionados* y 49.90% con 1.

- En las ventanas asimétricas con límite de tres vocablos a la izquierda y derecha, el mejor resultado fue 53.52% usando dos *términos relacionados*.

– En las ventanas asimétricas con límite de cuatro vocablos a la izquierda y derecha, el mejor resultado fue 51.72% usando 1 *término relacionado*.

- Como se observa en las tablas 10 y 11 las filas representan el número de *términos relacionados* y las columnas el tipo de contexto usado en la selección de dichos vocablos. Si tomamos en cuenta el comportamiento de dichos vocablos en los diferentes tipos de contexto (el promedio horizontal), el mejor resultado promedio se obtiene cuando se usa un solo término relacionado. Esta tendencia nos permite afirmar que mientras menor sea la cantidad de *términos relacionados*, se obtienen mejores resultados.

- Como ya se mencionó anteriormente Mc. Carthy reportó 48% de *precision* usando *English all-words* SENSEVAL-2 como corpus de evaluación y tomando en cuenta la totalidad de vocablos relacionados que proporciona por el tesauro de Lin. En términos generales, nuestro mejor resultado fue 53.52% usando sólo dos vocablos.

Analizando el tipo de contexto que se usa para seleccionar *términos relacionados*, se puede afirmar que:

– En las ventanas simétricas, el mejor contexto fue el 3-3, donde se obtuvo el resultado de 50.13%.

– En las ventanas asimétricas con límite de tres vocablos a la izquierda y derecha, el mejor contexto fue de cero vocablos a la izquierda y tres vocablos a la derecha. Con este contexto se obtuvo los mejores resultados de este experimento: 53.52%.

- Tomando en cuenta el comportamiento del contexto (promedio por columna) con respecto a cantidad de *términos relacionados*, las ventanas simétricas reportan que el contexto de tres vocablos a la izquierda y tres a la derecha, que reportó 46.65%, es el más óptimo para determinar *términos relacionados* para una instancia ambigua. Asimismo, el peor contexto es el sintáctico que reportó 40.65%. Creemos que esto se debe a los pocos niveles o ramas sintácticas del árbol que se han utilizado, lo que hace que el número de vocablos por contexto sean pocos.

- Al determinar que 3-3 es el tipo de contexto que proporciona mejores resultados, decidimos explorar las ventanas asimétricas dentro de dicho rango (0-3, 1-3, 2-3, 3-0, 3-1, 3-2), con la finalidad de determinar el lado del contexto que tiene mayor incidencia en la obtención de *términos relacionados*. El mejor resultado, 53.52%, se obtuvo usando una ventana de cero palabras a la izquierda y tres a la derecha. El mejor rendimiento del contexto del lado derecho parece confirmado por el promedio de los resultados de ambos lados de contexto: 51.42% en la derecha y 50.22% en la izquierda (ver tabla 14).

5.4.2 Lin

El método que se ha planteado para la extracción de una *lista ponderada* usando el tesauro de Lin como origen de información es el siguiente.

1. Obtener el contexto del vocablo ambiguo, el cual está conformado por un conjunto de vocablos.

$$C(w) = \{c_1, c_2, \ldots, c_i\},$$

Donde *w* es la instancia ambigua y c_i es cada uno de los vocablos que pertenecen a su contexto.

2. Obtener el conjunto de vocablos relacionados semánticamente con la palabra ambigua, los cuales son proporcionados por el tesauro de Lin, que a diferencia del tesauro de Mobby, cuantifica dicha relación.

$$S(w) = \{(sw_1, p_1), (sw_2, p_2), \ldots, (sw_i, p_i)\},$$

Donde *w* es el vocablo ambiguo, sw_i cada vocablo similar que proporciona el tesauro de Lin y p_i es el peso que se le asigna a la relación semántica entre sw_i y *w*.

3. Construir el modelo de espacio de palabras. Al igual que el experimento anterior las columnas representan cada uno de los miembros del contexto y las filas los vocablos similares. Asimismo, se incluye *w* como una fila más del WSM, asignando la máxima

similitud semántica entre *w* y cada uno de los miembros de su contexto. El peso que cuantifica la relación semántica entre c_i y sw_i se describe en el siguiente punto. En la tabla 12, el signo de interrogación indica que existe un proceso particular para calcular dicho peso.

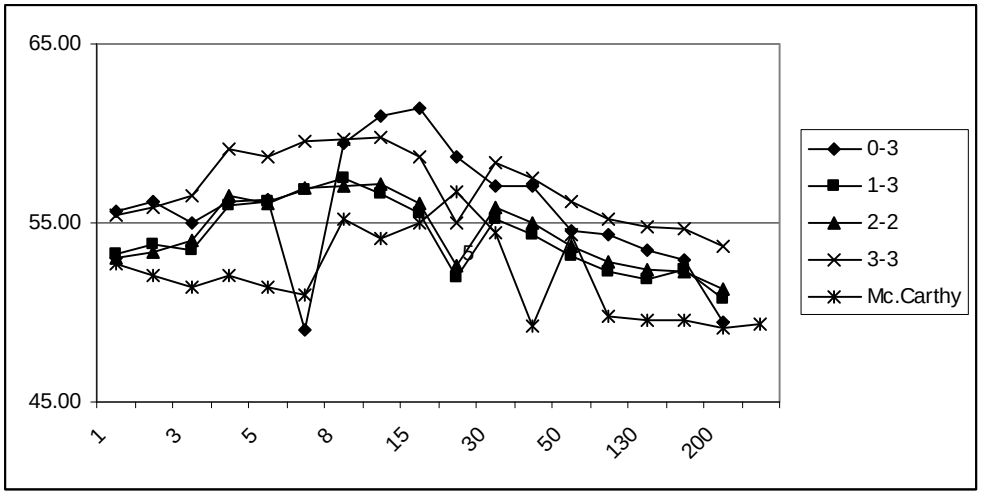
	c_1	c_2	c_3	c_i
sw_1	?	?	?	?
sw_2	?	?	?	?
sw_3	?	?	?	?
sw_j	?	?	?	?
w	1	1	1	1

Tabla 12 *WSM final Lin.*

4. Cuantificar la similitud semántica entre c_i y sw_i. Si bien es cierto que el tesauro de Lin proporciona un valor que determina la afinidad semántica entre dos vocablos, también suministra dicho valor para dos expresiones (conformadas por más de un vocablo), por ejemplo *more than million*, *about tour*, etc. Por ende es que se propuso un método alterno para cuantificar dicha similitud.
   - Si c_i se relaciona directamente con sw_j se toma el valor proporcionado por el tesauro de Lin.
   - Si c_i no se relaciona directamente con sw_j, se busca todas las expresiones que contenga a c_i y a sw_j, y se promedian los pesos que proporciona por el tesauro.
   - Si no existe ninguna expresión en la que co-ocurra c_i y sw_j, entonces el valor de esta relación se cuantifica con 0.
5. Obtenemos los *términos relacionados* de *w*, usando el esquema vectorial descrito en el experimento anterior.
6. Aplicar el algoritmo de maximización, utilizando el conjunto de *términos relacionados* obtenidos en el punto anterior

La gráfica 7 y tabla 13 muestran los resultados obtenidos cuando se usa como contexto de la instancia ambigua las ventanas de: 0-3, 1-3, 2-2, 2-3 para seleccionar los

elementos de la *lista ponderada*, siguiendo el método de WSD propuesto. Dichos resultados se comparan con el método de Mc. Carthy *et al*, que utiliza todos los vocablos suministrados por el tesauro de Lin.



Gráfica 7 *Experimento usando tesauro de Lin como origen de información.*

Términos relacionados	0-3	1-3	2-2	3-3	Mc. Carthy
1	56.00	53.65	54.23	56.71	52.75
2	55.69	53.24	53.04	55.46	52.06
3	56.15	53.84	53.39	55.83	51.37
4	55.04	53.46	54.03	56.50	52.06
5	56.21	55.94	56.54	59.13	51.37
6	56.29	56.23	56.10	58.67	51.03
8	48.97	56.86	56.92	59.52	**55.17**
10	59.47	**57.47**	57.06	59.67	54.13
15	60.94	56.58	**57.13**	**59.75**	55.01
20	**61.39**	55.51	56.11	58.68	56.74
30	58.66	52.00	52.56	54.97	54.44
40	57.08	55.24	55.84	58.39	49.28
50	57.08	54.38	54.96	57.48	54.31
100	54.61	53.12	53.69	56.15	49.81
130	54.38	52.25	52.82	55.23	49.60
160	53.47	51.83	52.39	54.78	49.55
200	52.91	52.36	52.32	54.71	49.18
Todos	49.46	50.77	51.32	53.66	49.31
Promedio	**59.05**	54.15	54.47	56.96	52.07

Tabla 13 *Resultados al usar el tesauro de Lin como origen de información.*

Los resultados obtenidos evidencian el comportamiento de experimentos anteriores, es decir, que es mejor utilizar un grupo reducido de vocablos que tomar un conjunto más amplio. Asimismo, al analizar los valores de la tabla, específicamente el promedio por columna, se puede concluir que cualquier ventana contextual obtiene mejores *términos relacionados* que los que usa el método de desambiguación planteado por Mc. Carhty *et al*. Finalmente, se puede concluir que la ventana contextual que permite obtener mejores *términos relacionados* (al ser procesados por el algoritmo de maximización) es el contexto de la derecha del vocablo ambiguo.

5.4.3 Google

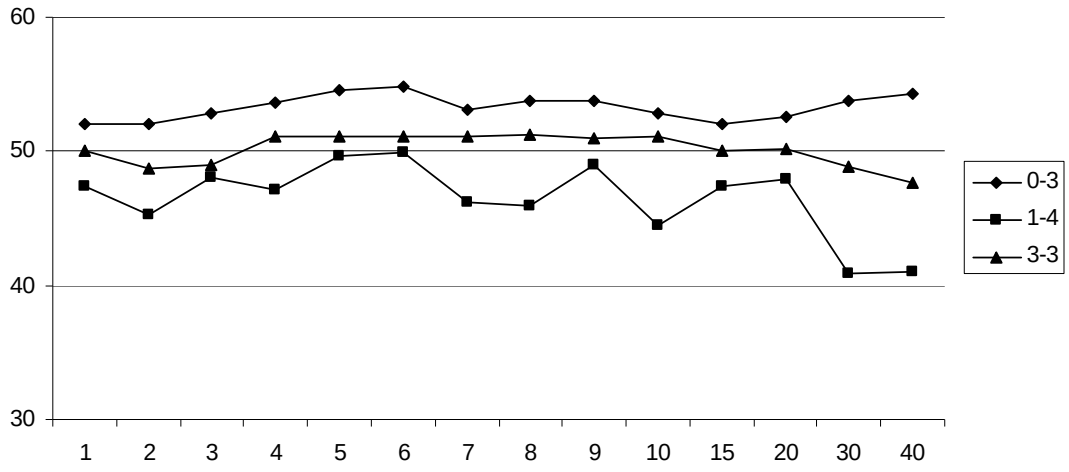
El corpus de Google es el repositorio más grande que existe hasta el momento y está conformado por *n-gramas*. En los experimentos que se presentan en esta sección se ha utilizado los bigramas de Google, para obtener *términos relacionados* con la instancia ambigua tomando en cuenta su contexto. Se han tomado diferentes tipos de amplitud en las ventanas contextuales, tal como se puede apreciar en la tabla 14.

Términos relacionados	0-3	1-4	3-3
1	52.09	47.40	50.01
2	52.09	45.32	48.71
3	52.83	48.08	49.02
4	53.58	47.09	51.09
5	54.56	49.65	51.08
6	**54.81**	**49.88**	**51.13**
7	53.08	46.21	51.11
8	53.82	45.88	51.18
9	53.82	48.98	51.02
10	52.83	44.43	51.06
15	52.09	47.40	50.08
20	52.59	47.86	50.22
30	53.82	40.87	48.90
40	54.32	41.03	47.68
Promedio	**53.31**	46.43	50.16

Tabla 14 *Resultados al usar el corpus de Google como origen de información.*

En la gráfica 8 y la tabla 14 se puede observar que los mejores resultados se consiguen con seis *términos relacionados*, usando cualquiera de los tres tipos de contextos

procesados, lo que confirma el comportamiento de los experimentos anteriores: A menor cantidad de *términos relacionados* los resultados del método de desambiguación incrementan Asimismo, tomando en cuenta el promedio de resultados, la ventana contextual con la que se obtiene mejores resultados es la que usa sólo tres vocablos a la derecha de la instancia ambigua.



Gráfica 8 *Experimento usando corpus de Google como origen de información.*

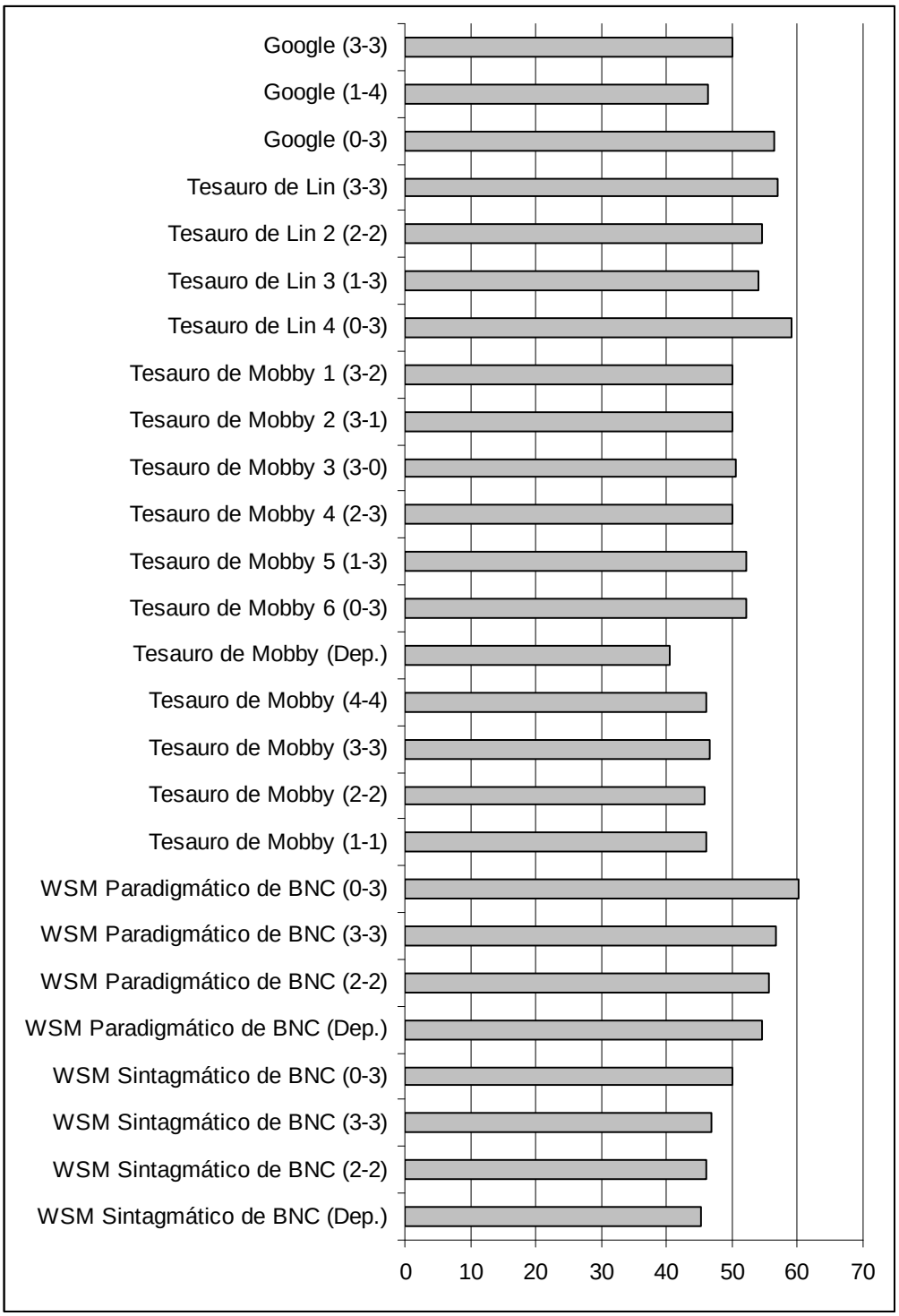
5.4.4 Resumen

Después de haber realizado múltiples experimentos usando diferentes orígenes de información, se presenta la tabla 15 y gráfica 9, en la que se cuantifica el rol de los *términos relacionados* que provee cada origen de información en el método de desambiguación planteado. En dicha tabla se incluye la columna *Tipo de contexto*, ya que éste es un parámetro que influye directamente en la obtención de los *términos relacionados*. Finalmente, la tercera columna hace referencia a la media de cada uno de los experimentos realizados en secciones anteriores.

En conclusión, se afirma que los Modelos de Espacio de Palabras que proporcionan *términos relacionados* paradigmáticamente tienen un mejor rendimiento en el método de desambiguación planteado.

Origen de información	Tipo de contexto	*Precision* evaluada con SENSEVAL-2 (%)
WSM Sintagmático de BNC	Dependencias	45.20
WSM Sintagmático de BNC	2-2	46.07
WSM Sintagmático de BNC	3-3	46.88
WSM Sintagmático de BNC	0-3	49.98
WSM Paradigmático de BNC	Dependencias	54.60
WSM Paradigmático de BNC	2-2	55.67
WSM Paradigmático de BNC	3-3	56.77
WSM Paradigmático de BNC	0-3	**60.02**
Tesauro de Mobby	1-1	45.96
Tesauro de Mobby	2-2	45.81
Tesauro de Mobby	3-3	46.65
Tesauro de Mobby	4-4	46.08
Tesauro de Mobby	Dependencias	40.35
Tesauro de Mobby	0-3	52.19
Tesauro de Mobby	1-3	52.08
Tesauro de Mobby	2-3	49.99
Tesauro de Mobby	3-0	50.55
Tesauro de Mobby	3-1	50.05
Tesauro de Mobby	3-2	50.05
Tesauro de Lin	0-3	59.05
Tesauro de Lin	1-3	54.15
Tesauro de Lin	2-2	54.47
Tesauro de Lin	3-3	56.96
Google	0-3	56.31
Google	1-4	46.43
Google	3-3	50.16

Tabla 15 *Comparación de orígenes de información.*



Gráfica 9 *Comparación de orígenes de información*

5.5 Robustez del método

En esta sección se presentan algunas características relevantes que afectan directamente a los resultados que proporciona el método planteado, tales como el número de *términos relacionados* que procesa el algoritmo de maximización y el tópico del corpus de entrenamiento.

5.5.1 Impacto de términos relacionados

En esta sección presentamos un análisis exhaustivo del impacto de los *términos relacionados* (*términos relacionados*) o *lista ponderada* en el algoritmo de maximización cuando éste se aplica a la detección del sentido predominante y a la desambiguación de sentidos de palabras. McCarthy *et al*., quien propuso dicho algoritmo para determinar el sentido predominante de un vocablo ambiguo utilizó los primeros 10, 30, 50 y 70 *términos relacionados* y concluyó que dicha cantidad no es una característica relevante que influya en el rendimiento de su método; por lo cual en todos sus experimentos sólo tomó 50, sin embajo, en Tejada *et al*., [63] el mejor resultado se obtuvo utilizando los primeros 10 *términos relacionados*. Asimismo, se notó una diferencia importante en los resultados cuando el número de *términos relacionados* varía.

McCarthy *et al*. determinó automáticamente los sentidos predominantes de SemCor y English all-words SENSEVAL-2, obteniendo una precisión de 54%. Usando esta heurística como método de desambiguación, la precisión que obtuvo fue de 48%. En la prueba *English all-words* de SENSEVAL-2, 25% de los sustantivos son monosémicos. Tomando en cuenta esta característica (permitida en dicho concurso), el resultado fue de 64%, que como se puede apreciar en la tabla 4 (al inicio de este capítulo) es mejor que cualquier otro método no supervisado, los cuales sí toman en cuenta el contexto del vocablo ambiguo.

Asimismo, utilizó los sentidos predominantes existentes en SemCor para determinar los de *English all-words* SENSEVAL-2, obteniendo una precisión de 69%. Después, usó los sentidos predominantes de *English all-words* SENSEVAL-2 para determinar los del

mismo corpus de evaluación, obteniendo una precisión de 92%. Con estos resultados, demostró claramente que la heurística del sentido predominante es realmente poderosa.

El rol del algoritmo de maximización siempre es el mismo: Asignar un sentido a la instancia ambigua tomando en cuenta cada uno de los vocablos de la *lista ponderada* (ya sea dinámica o estática). El sentido que elige dicho algoritmo está condicionado no sólo por el número de *términos relacionados* que procesa, sino por la calidad de la relación semántica que existe entre cada uno de los miembros de la *lista* y la instancia ambigua. Para aislar este efecto, el cual depende del recurso léxico que se usa como proveedor de dichos términos, se ha utilizado el mismo recurso que fue empleado por McCarthy *et. al*. en sus experimentos: El tesauro de Lin.

En los resultados obtenidos en esta sección se determinó que el sentido predominante de casi la mitad de los sustantivos de SENSEVAL-2, 41.18%, siempre es exitoso, es decir, no importa el número *términos relacionados* procesados por el algoritmo de maximización, mientras que 26.47% siempre lo hace erróneamente, y sólo 32.35% se ven afectados por la cantidad de *términos relacionados* procesados. Los experimentos se han dividido en dos:

- Impacto de *términos relacionados* en la detección del sentido predominante.
- Impacto de *términos relacionados* en Word Sense Disambiguation.

a. Impacto en la detección del sentido predominante

Para determinar el impacto de este algoritmo en la detección del sentido predominante, se usa como corpus de evaluación los sentidos predominantes de los sustantivos del corpus *English all-words* SENSEVAL-2, cuya instancia ambigua haya ocurrido al menos dos veces, obteniendo un total de 34 sustantivos polisémicos, donde cada uno presenta al menos dos sentidos (ver tabla 16). Asimismo, se ha comparado los sentidos predominantes que el algoritmo de maximización elige para cada uno de los 34 sustantivos con los que estipulados en WordNet.

Para cada uno de los vocablos listados en la tabla 16, se aplicó el algoritmo de

maximización, tomando los primeros 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 40, 50, 100,120, 150, 200, 230, 260, 300, 330, 360 y 400 *términos relacionados* suministrados por el tesauro de Lin, obteniendo los siguientes resultados:

Vocablo	Sentido Predominante	Nro. de Sentidos	Vocablo	Sentido Predominante	Nro. de Sentidos
church	2	4	individual	1	2
field	4	13	child	2	4
bell	1	10	risk	1	4
rope	1	2	eye	1	5
band	2	12	research	1	2
ringer	1	4	team	1	2
tower	1	3	version	2	6
group	1	3	copy	2	3
year	1	4	loss	5	8
vicar	3	3	colon	1	5
sort	2	4	leader	1	2
country	2	5	discovery	1	4
woman	1	4	education	1	6
cancer	1	5	performance	2	5
cell	2	7	school	1	7
type	1	6	pupil	1	3
growth	1	7	student	1	2

Tabla 16 *Sustantivos del corpus English all-words SENSEVAL-2 usados como gold evaluation corpus.*

- Los sentidos predominantes de los vocablos: *rope*, *tower*, *vicar*, *woman*, *cancer*, *cell*, *type*, *individual*, *research*, *team*, *copy*, *colon*, *leader* y *discovery*, siempre se determinaron correctamente, sin importar la cantidad de *términos relacionados* utilizados en el proceso. (ver tabla 16)

- Los sentidos predominantes de los vocablos: *band*, *ringer*, *year*, *sort*, *child*, *version*, *loss*, *performance* y *school* siempre se determinaron incorrectamente, sin importar la cantidad de *términos relacionados* utilizados en el proceso (ver tabla 17).

- El éxito en la detección de los sentidos predominantes de los 11 restantes: *church*, *field*, *bell*, *group*, *country*, *growth*, *risk*, *eye*, *education*, *pupil*, *student* dependen del número de *términos relacionados*, tal como se puede apreciar en la gráfica 10 y en la tabla 19.

Los resultados sólo muestran 200 *términos relacionados*, aunque hicimos experimentos con 230, 260, 300, 330, y 360, sin embargo sólo los vocablos *group*, *country*, *growth*, *risk*, *eye*, *education*, *student* presentan *términos* mayores a 200 (ver gráfica 11)

Sentido predominante Correcto		Sentido predominante Incorrecto	
Vocablo ambiguo	Nº. Sentidos	Vocablo ambiguo	Nº. Sentidos
rope	2	band	12
tower	3	ringer	4
vicar	3	year	4
woman	4	sort	4
cancer	5	child	2
cell	7	version	2
type	6	loss	8
individual	1	performance	5
research	2	school	7
team	2		
copy	3		
colon	5		
leader	2		
discovery	4		
Promedio	3.50		5.33

Tabla 17 *Número de sentidos de vocablos que siempre se desambiguan exitosamente.*

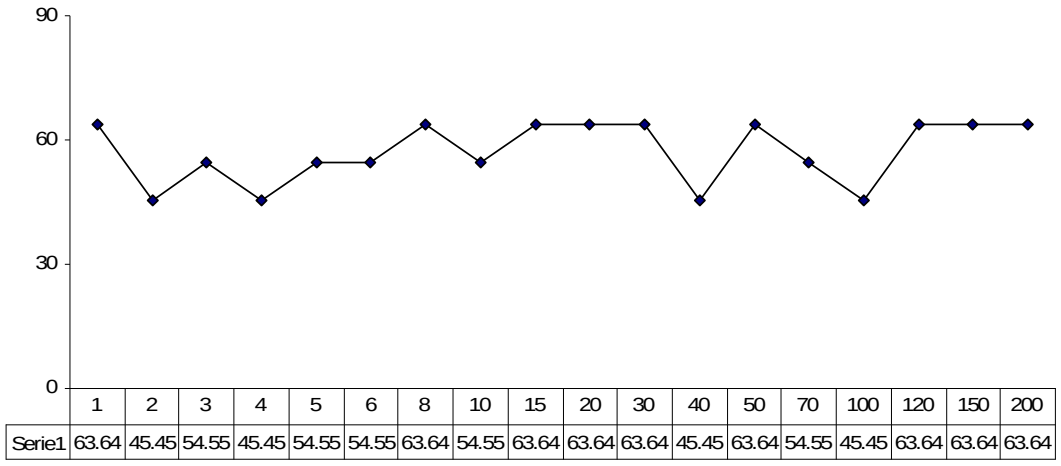
El promedio del número de sentidos de sustantivos cuyo sentido predominante siempre se determina exitosamente es menor que el de aquellos cuyo sentido predominante siempre se determina sin éxito (ver tabla 18).

	Nº Sustantivos	(%) Porcentaje
Total de sustantivos evaluados	34	100.00
Aciertos exitosos sin importar el Nº de *términos relacionado*.	14	41.18
Aciertos fracasados sin importar el Nº de *términos relacionados*.	9	26.47
Sustantivos cuyo resultado depende del Nº de *términos relacionados*.	11	32.35

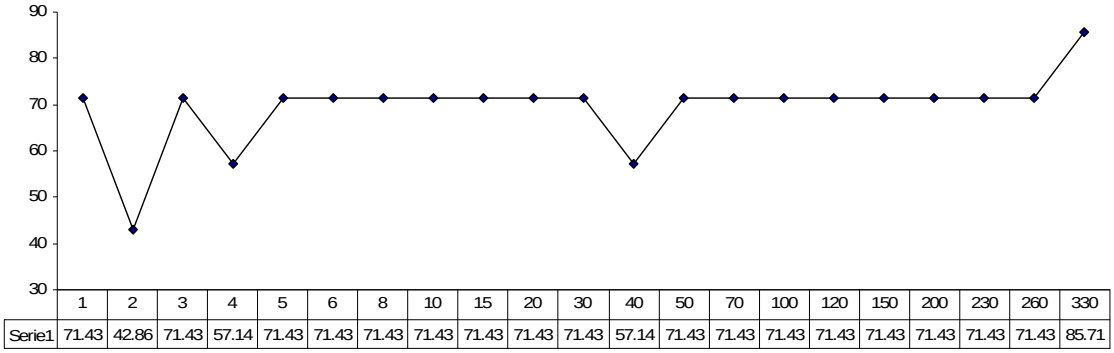
Tabla 18 *Estadística general en la detección del sentido predominante.*

	1	2	3	4	5	6	8	10	15	20	30	40	50	70	100	120	150	200
church	✓	×	×	×	×	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
field	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓	×
bell	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	×	×	×	✓
group	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
country	✓	✓	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
growth	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
risk	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
eye	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
education	✓	✓	✓	×	×	×	×	×	×	×	×	×	✓	✓	✓	✓	✓	✓
pupil	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
student	✓	×	×	×	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×

Tabla 19 *Términos relacionados en la detección del sentido predominante.*



Gráfica 10 *Sentido predominante de sustantivos con términos relacionados menores iguales a 200.*



	1	2	3	4	5	6	8	10	15	20	30	40	50	70	100	120	150	200	230	260	330
Serie1	71.43	42.86	71.43	57.14	71.43	71.43	71.43	71.43	71.43	71.43	71.43	57.14	71.43	71.43	71.43	71.43	71.43	71.43	71.43	71.43	85.71

Gráfica 11 *Sentido predominante de sustantivos con términos relacionados menores iguales a 330.*

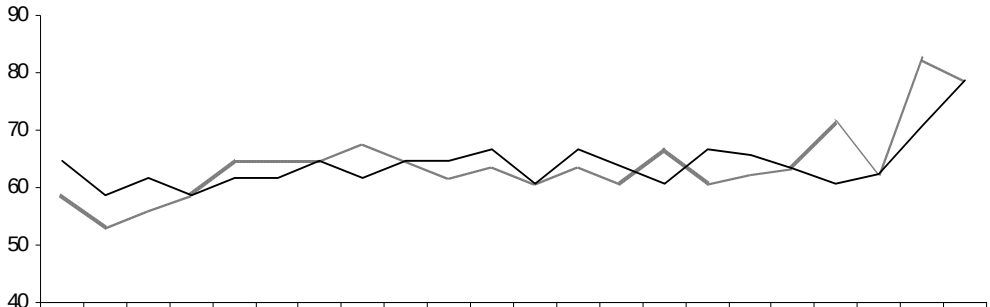
De los 34 sustantivos evaluados, 23 son *inmunes* al número de *términos relacionados* utilizados por el algoritmo de maximización, es decir, que 14 de ellos siempre se desambiguan bien y 9 siempre lo hacen mal. Los 11 restantes sí se ven afectados por el número de *términos relacionados*, tal como lo muestran las gráficas 11 y 12. En la primera de ellas, la precision varía entre 45.45% y 63.64% mientras que en la segunda varía entre 42.85% y 85.71%.

La irregularidad de las gráfica 10 y 11, y los resultados mostrados en la tabla 19 indican que el rol del número de *términos relacionados* en el algoritmo de maximización aplicado a la obtención del sentido predominante no es tan decisivo como plantea el trabajo de McCarthy *et al*. [36], ya que sólo 32.5% de los sustantivos son influenciados por el número de *términos relacionados* proporcionados por el tesauro de Lin.

A pesar de la irregularidad de la gráfica 10, se podría afirmar que los mejores resultados se estabilizan con 15, 20 y 30 *términos relacionados*. Asimismo, los resultados tienden a mejorar cuando superan los 100 *términos*. Cuando se utiliza sólo un *término,* los resultados son similares a los obtenidos con más de 100, lo cual sólo puede ser explicado por chance, ya que un solo vocablo, teóricamente no proporciona suficiente evidencia para que el algoritmo de maximización pueda tomar una decisión coherente. La gráfica 11 es más ilustrativa. En ella, se confirma lo mencionado, y además se puede apreciar que

mientras los *términos relacionados* se incrementan, la precisión se estabiliza y mejora.

La gráfica 12 compara el sentido predominante que elige el algoritmo de maximización para los 34 sustantivos de la tabla 16, con los estipulados en el corpus *English all-words* de SENSEVAL-2 y el corpus Semcor. El comportamiento de ésta confirma los resultados anteriores: Entre 15 y 50 *términos relacionados* los resultados se estabilizan y, cuando se supera los 100 *términos* los resultados tienden a mejorar.



	1	2	3	4	5	6	8	10	15	20	30	40	50	70	100	120	150	200	230	260	330	360
SENSEVAL-2	58.8	52.9	55.8	58.8	64.7	64.7	64.7	67.6	64.7	61.7	63.6	60.6	63.6	60.6	66.6	60.6	62.5	63.3	71.4	62.5	82.3	78.5
SEMCOR	64.7	58.8	61.7	58.8	61.7	61.7	64.7	61.7	64.7	64.7	66.6	60.6	66.6	63.6	60.6	66.6	65.6	63.3	60.7	62.5	70.5	78.5

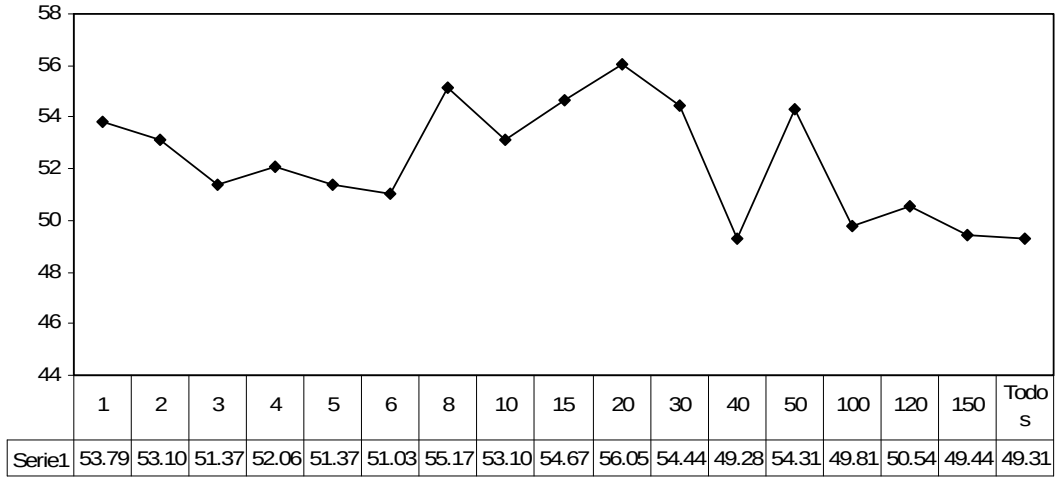
Gráfica 12 *Algoritmo de maximización aplicado a la detección del sentido predominante cuando se evalúa con SENSEVAL-2 y Semcor.*

b. Impacto en WSD

Para determinar el impacto de este algoritmo en la desambiguación de sentidos de palabras, utilizamos el pequeño corpus *English all-words* de SENSEVAL-2 como corpus de evaluación. Aunque el objetivo para el que fue creado este algoritmo era para obtener sentidos predominantes usando sólo texto como origen de información, también es posible aplicarlo a la desambiguación de sentidos de palabras como método de respaldo (en inglés *backoff*).

En estos experimentos no tomamos en cuenta el contexto del vocablo ambiguo, simplemente aplicamos la heurística del sentido predominante a WSD. Esto implica que el sentido que expresa un vocablo desambiguado exitosamente es el predominante. La gráfica

13 muestra los resultados obtenidos.



	1	2	3	4	5	6	8	10	15	20	30	40	50	100	120	150	Todos
Serie1	53.79	53.10	51.37	52.06	51.37	51.03	55.17	53.10	54.67	56.05	54.44	49.28	54.31	49.81	50.54	49.44	49.31

Gráfica 13 *Algoritmo de maximización aplicado a WSD.*

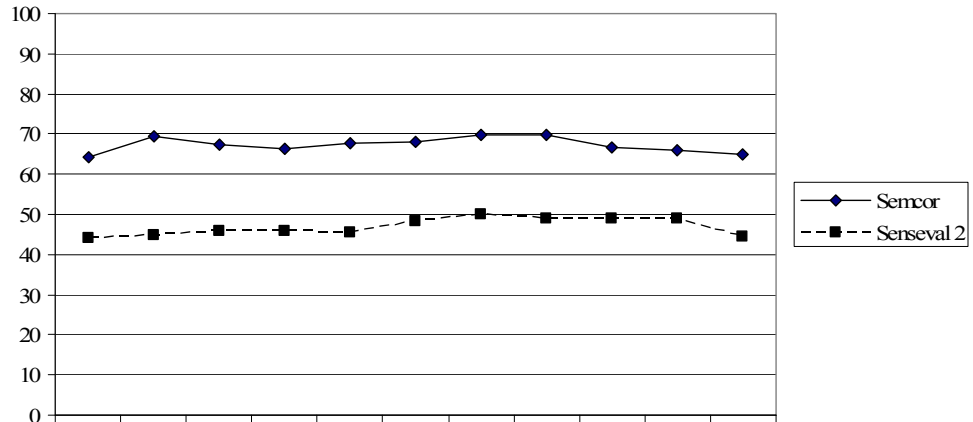
El mejor resultado fue de 56.74% cuando se utilizaron 20 *términos relacionados*. El peor fue de 49.28% cuando se utilizaron 40 *términos* y el segundo peor resultado fue 49.31% cuando se procesaron *todos* los *términos* (esta cantidad varía por vocablo entre 80 y 1000). Los mejores resultados tienden a ocurrir entre 8 y 50 *términos relacionados*. Es necesario esclarecer que mientras el número de *términos relacionados* incrementa, los resultados declinan, mientras que en la detección del sentido predominante sucede lo contrario.

5.5.2 Impacto del corpus de entrenamiento

La gráfica 14 muestra la *precision* obtenida cuando se construye el WSM paradigmático con el 90% de SemCor y se evalúa con 10% de Semcor y la totalidad de SENSEVAL-2.

Los resultados de la gráfica 14 demuestran que construir un WSM con el mismo corpus con el que se evalúa es más beneficioso que entrenar y evaluar con diferentes corpus. En la tabla 5, el mejor resultado, 73.07%, se obtuvo cuando se entrenó con British National Corpus y se evaluó con Semcor; sin embargo el segundo mejor resultado, 69.86%,

se obtuvo cuando se entrenó con el 90% de Semcor y se evalúo con el 10% restante.



	10	20	30	40	50	60	70	100	200	500	1000
Semcor	64.23	69.44	67.36	66.43	67.8	68.15	69.86	69.86	66.75	65.89	65.06
Senseval 2	44.22	44.77	45.91	45.76	45.55	48.12	49.84	48.8	49.05	49.1	44.55

Gráfica 14 *Precisión, cuando se entrenó con el 90% de SemCor y se evaluó con 10% de SemCor y SENSEVAL-2.*

Si tomamos en cuenta las dimensiones de ambos corpus de entrenamiento (BNC es 100 veces más grande que Semcor), y la hipótesis tradicional de que mientras más grande sea el corpus de entrenamiento, la calidad semántica de los vocablos que proporciona un modelo de espacio de palabras es mejor, podemos justificar la consecución del mejor resultado cuando se entrena con BNC.

A pesar de tales características, la diferencia con el segundo mejor resultado no supera los tres puntos porcentuales (69.86 vs. 73.07). En este caso específico, se entrenó con el 90% de SemCor y se evaluó con el 10% restante. Basándonos en el principio de Yarowsky [67] (un sentido por documento), concluimos que construir un WSM sobre una temática específica es un buen método alterno (sobretodo si el proceso de construcción es más rápido). Dicho principio, se puede confirmar al observar que el peor resultado, 41.05% se obtuvo cuando se entrenó con Semcor y se evaluó con SENSEVAL-2. Además, el promedio de los resultados que se obtiene cuando se entrena y evalúa con el mismo corpus es mayor que el obtenido al entrenar con SemCor (61.73 vs. 58.38).

Capítulo 6
Conclusiones y trabajo futuro

6.1 Conclusiones

- Los modelos de espacio de palabras paradigmáticos tienen mayor rendimiento que los sintagmáticos, cuando los *términos relacionados* que ambos proporcionan se utilizan en el método de desambiguación de sentidos de palabras planteado en esta tesis.

- La idea de usar el mismo corpus de evaluación como corpus de entrenamiento en la construcción de un WSM es una buena alternativa, que puede aplicarse a diversas tareas del procesamiento de lenguaje natural, en las que se requiera desambiguar vocablos que pertenecen a la misma temática, tal como lo demuestran los resultados obtenidos al entrenar con el 90% de Semcor y evaluar con el 10% restante.

- El mejor origen de información para el método de desambiguación planteado son los modelos de espacio de palabras paradigmáticos, seguido por el tesauro de Lin y el corpus de Google (ambos construidos automáticamente).

- Los *términos relacionados* que proporciona el tesauro manual de Mobby son los que obtuvieron menor rendimiento al ser utilizados por el algoritmo de maximización cuando éste se aplica a la desambiguación de sentidos de palabras.

- El método de desambiguación proporciona mejores resultados cuando se toman sólo los términos más relacionados con la instancia ambigua. A mayor incremento de estos, los resultados tienden a bajar. Este comportamiento se observó en todas las fuentes de información empleadas en la recuperación de términos relacionados.

- De todos los contextos utilizados en el proceso de recuperación de *términos relacionados*, las dependencias sintácticas obtuvieron los resultados más bajos en el método de desambiguación planteado. Una de las posibles causas de este resultado, es

que sólo se tomaron el nivel inmediato anterior y superior del vocablo ambiguo en el árbol de dependencias sintácticas.

- El tipo de contexto del vocablo ambiguo condiciona sus *términos relacionados*, los cuales inciden directamente en el etiquetado semántico de la instancia ambigua. Pese a haber analizado ventanas contextuales simétricas, asimétricas y de dependencias sintácticas, no se puede afirmar que alguno de estos sea el más óptimo, ya que en todos los experimentos no se observó el mismo patrón de comportamiento, como en el caso de las dependencias sintácticas; sin embargo en la mayoría de los casos el contexto ubicado a la derecha de la instancia ambigua es el que tiene mayor influencia en la selección de términos relacionados.

- El número de *términos relacionados* no es un factor que condicione el éxito del algoritmo de maximización cuando se aplica a la detección del sentido predominante. Como prueba de ello, casi la mitad de los sustantivos de SENSEVAL-2, 41.18%, siempre se determina exitosamente sin importar el número de *términos relacionados*, 26.47% siempre lo hace erróneamente. Sólo una tercer parte, 32.35%, se ven afectados por la cantidad de *términos relacionados* procesados

6.2 Trabajo futuro

- Elaborar un modelo de espacio de palabras mixto que permita almacenar y procesar ambos tipos de relaciones en el lenguaje: paradigmática y sintagmática.

- Aplicar los *términos relacionados* que proporciona el modelo de espacio de palabras a otras tareas del procesamiento de lenguaje natural, tales como recuperación de información, traducción automática, etc.

- Crear modelos de espacio de palabras, que permita relacionar palabras (filas) y n-gramas (columnas), específicamente bigramas y trigramas.

- Aplicar los *términos relacionados* que proporcionan los diferentes orígenes de

información a otros algoritmos de desambiguación.

- Integrar recursos léxicos creados manualmente con el modelo de espacio de palabras construido automáticamente.

Glosario

Ambigüedad. Término que hace referencia a aquellas estructuras gramaticales que pueden entenderse de varios modos o admitir distintas interpretaciones y dar, por consiguiente, motivo a dudas, incertidumbre o confusión.

Ambigüedad léxica. La ambigüedad léxica es aquella que se presenta en la categoría gramatical de un vocablo. Es decir, un vocablo puede tener más de un rol gramatical en diferentes contextos.

Ambigüedad semántica. La ambigüedad semántica es aquella que se presenta en una estructura gramatical, de tal manera que ésta puede expresar diferentes sentidos dependiendo del contexto local, el tópico global y el mundo pragmático en el que se manifiesta.

Ambigüedad sintáctica. La ambigüedad sintáctica, también conocida como estructural, es aquella que se presenta en oraciones, de tal manera que éstas puedan ser representadas por más de una estructura sintáctica.

Analizador sintáctico. Un analizador sintáctico para un lenguaje natural, es un programa que construye árboles de estructura de frase o de derivación para las oraciones de dicho lenguaje, además de proporcionar un análisis gramatical, separando los términos en constituyentes y etiquetando cada uno de ellos. Asimismo, puede proporcionar información adicional acerca de las clases semánticas (persona, género) de cada palabra y también la clase funcional (sujeto, objeto directo, etc.) de los constituyentes de la oración.

Antonimia. La antonimia es una relación entre dos palabras que expresan ideas opuestas o contrarias. Por ejemplo, los vocablos virtud y vicio; claro y oscuro; antes y después.

Aprendizaje supervisado. El aprendizaje supervisado se asemeja al método de enseñanza tradicional con un profesor que indica y corrige los errores del alumno hasta que éste aprende la lección. En el caso de la desambiguación supervisada se entrena un

clasificador usando un corpus de texto etiquetado semánticamente para obtener el contexto en el que usualmente se presenta cada sentido del vocablo ambiguo. Este clasificador desambiguará solo aquellos vocablos y sentidos que hayan participado en el entrenamiento previo.

Aprendizaje no supervisado. En el aprendizaje no supervisado no existe un profesor que corrija los errores al alumno; ya que éste recuerda más al autoaprendizaje. El alumno dispone del material de estudio; pero nadie lo controla. En el caso de la desambiguación no supervisada también es posible usar un corpus de texto etiquetado como fase entrenamiento; sin embargo los algoritmos no supervisados generalizan esta información para cualquier vocablo ambiguo; aunque no haya estado presente en el corpus de entrenamiento.

Árbol de constituyentes. Un árbol de constituyentes es una estructura de datos que perite categorizar una oración en sus partes de oración. En el llamado sistema o método de constituyentes la principal operación lógica es la inclusión de elementos en conjuntos, así éstos pertenecen a una oración o a una categoría. Según esta aproximación, una oración es segmentada en constituyentes, cada uno de los cuales es consecuentemente segmentado. Así, esto favorece un punto de vista analítico.

Árbol de dependencias. Un árbol de dependencias es una estructura de datos que perite obtener las relaciones de dependencia sintáctica entre un núcleo y conjunto de modificadores. La aproximación de dependencias se centra en las relaciones entre las unidades sintácticas últimas, es decir, en las palabras. La principal operación aquí consiste en establecer relaciones binarias. Según esta idea, una oración se construye de palabras, unidas por dependencias.

Biblioteconomía. La biblioteconomía es la disciplina encargada de la conservación, organización y administración de las bibliotecas, incluso cuando son digitales. Esta disciplina es utilizada por los sistemas de recuperación de información.

Bootstrapping. *Bootstrapping* es un término inglés que se refiere al proceso mediante el cual se han desarrollado o implementado soluciones cada vez más complejas a

partir de otras más simples. El entorno más simple sería, quizás, un editor de textos muy sencillo y un programa ensamblador. Utilizando estas herramientas, se puede escribir un editor de texto más complejo y un compilador simple para un lenguaje de más alto nivel y así sucesivamente, hasta obtener un entorno integrado de desarrollo y un lenguaje de programación de muy alto nivel.

Cabeza. *Cabeza* es un término utilizado en este trabajo para referenciar al vocablo que gobierna una relación de dependencia sintáctica, de tal manera que se puede obtener muchos modificadores sintácticos para una misma *cabeza*.

Categoría gramatical. El término categoría gramatical o parte de la oración, que en inglés se denomina POS (*part of speech*) es una clasificación de las palabras de acuerdo a la función que desempeñan en la oración. La gramática tradicional distingue nueve categorías gramaticales: sustantivo, determinante, adjetivo, pronombre, preposición, conjunción, verbo, adverbio, interjección. No obstante, para algunos lingüistas, las categorías gramaticales son una forma de clasificar ciertos rasgos gramaticales, como por ejemplo: modo, aspecto, tiempo y voz.

Colocación gramatical. Una colocación gramatical es un conjunto de dos o más palabras las cuales expresan una idea específica. El significado que expresa cada término de una colocación difiere de la semántica que dichos términos proporcionan cuando se usan de manera conjunta.

Contexto local. El contexto local de un vocablo, que también es conocido como microcontexto, engloba a un conjunto de palabras cercanas a dicho vocablo. Esta cercanía puede estar limitada por una vecindad de palabras coocurrentes, por la oración en la cual se encuentra dicho vocablo o incluso, por el árbol sintáctico al cual pertenecen.

Dependencia convencional. Una relación de dependencia sintáctica *convencional* es aquella en la que una pareja de vocablos mantiene una relación de dependencia tradicional especificada por el árbol de dependencias sintácticas. Más explícitamente, las flechas que salen de un vocablo hacia otros se consideran los modificadores sintácticos del primero

Dependencia sin preposición. Una relación de dependencia especial es aquella que excluye la preposición cuando ésta se presenta como modificador del núcleo de una relación, de tal manera que el término que depende de ésta pasa a ser el modificador del núcleo.

Dependencia especial. Una relación de dependencia sintáctica *especial* incluye como parte del conjunto de los modificadores del núcleo de la relación, el vocablo al que modifica el mismo núcleo. De esta manera, incrementa su número de modificadores.

Diccionario computacional. Un diccionario computacional surge al convertir un diccionario normal, creado exclusivamente para el uso humano, a formato electrónico. Estos diccionarios proveen información sobre sentidos de vocablos ambiguos, lo cual puede ser explotado por el área de desambiguación de sentidos de palabras.

Dominio. El término dominio hace referencia a la temática general que expresa un documento o texto en su totalidad.

Esquema TF-IDF. El esquema TF-IDF o modelo de espacio vectorial es una arquitectura que generalmente se aplica a tareas de clasificación y similitud de documentos. Ésta representa un documento con un vector y, cuando se desea comparar dos documentos, se compara dos vectores multidimensionales.

Etiqueta semántica. Este término se utiliza para hacer referencia a un sentido específico de un vocablo ambiguo, tomando en cuenta alguna fuente de información. Por ejemplo, WordNet.

Herramientas lingüísticas. Esta expresión hace referencia a diversos programas o aplicaciones usados en el procesamiento de lenguaje natural, tales como analizadores sintácticos, morfológicos, diccionarios electrónicos, ontologías, entre otros.

Hiperónimo. Un hiperónimo es una palabra cuyo significado incluye al de otra u otras. Por ejemplo, pájaro respecto a jilguero y gorrión

Hipónimo. Un homónimo es una palabra cuyo significado está incluido en el de otra. Po ejemplo, gorrión respecto a pájaro.

Recuperación de información. La recuperación de información es la ciencia encargada de buscar información en archivos de diversos tipos, en meta-datos y en bases de datos textuales, de imágenes o de sonidos. La plataforma sobre la cual es posible realizar dichas búsquedas se extiende desde computadoras de escritorio, redes de computadoras privadas o públicas hasta intranets e internet

Lema. Lema es el vocablo más representativo de un conjunto de palabras que comparten cierta morfología común. Por ejemplo el lema de los vocablos perrito, perraso, perra; es perro.

Lingüística computacional. La lingüística computacional puede considerarse una disciplina de la lingüística aplicada y la inteligencia artificial. Tiene como objetivo la creación e implementación de programas computacionales que permitan la comunicación entre el hombre y la computadora, ya sea mediante texto o voz.

Macrocontexto. El macrocontexto está conformado por palabras de gran contenido semántico (sustantivos, adjetivos y verbos), las cuales coocurren con un sentido específico de un vocablo ambiguo, usando varias oraciones como fuente de información.

Meronimia. La meronimia es la relación semántica entre una unidad léxica que denota una *parte* y lo que denota el correspondiente *todo*.

Metonimia. La metonimia consiste en designar algo con el nombre de otra cosa tomando el efecto por la causa o viceversa, el autor por sus obras, el signo por la cosa significada, etc. Por ejemplo, las canas por la vejez, leer a Virgilio, por leer las obras de Virgilio; el laurel por la gloria, etc.

Peso de similitud. El peso de similitud entre dos definiciones es un valor calculado por alguna medida de similitud o relación semántica. Este peso refleja cuan similares o parecidas pueden ser dos palabras, basándose en la definición de cada una.

Polisemia. Polisemia a la capacidad que tiene una sola palabra para expresar muy distintos significados. Pluralidad de significados de una palabra o de cualquier signo lingüístico y de un mensaje, con independencia de la naturaleza de los signos que lo constituyen.

Sentido predominante. El sentido predominante de un vocablo es aquel que generalmente suele ser el más usado por los hablantes de una lengua específica.

Recurso sintáctico. Un recurso sintáctico, en este trabajo, es una base de datos que almacena relaciones de dependencia sintácticas, las cuales pueden ser extraídas desde cualquier corpus de texto.

Sinonimia. La sinonimia es una relación de semejanza de significados entre determinadas palabras (llamadas sinónimos) u oraciones. También consiste en usar voces sinónimas de significación similar para amplificar o reforzar la expresión de un concepto.

Synset. *Synset* es un término usado en WordNet, donde hace referencia a un conjunto de sinónimos, comprendidos por palabras o colocaciones. Dicho conjunto se encuentra conectado con otros *synsets* mediante relaciones jerárquicas existentes en WordNet.

Tesauro. El tesauro es un sistema que organiza el conocimiento basado en conceptos que muestran relaciones entre vocablos. Las relaciones expresadas comúnmente incluyen jerarquía, equivalencia y asociación (o relación). Los tesauros también proporcionan información como sinonimia, antonimia, homonimia, etc.

Tripletas de dependencia. Una tripleta de dependencia sintáctica esta conformada por dos vocablos unidos bajo cierta relación de dependencia. En este trabajo también se le denomina tupla sintáctica.

Troponimia. La troponimia es una relación que se da entre verbos. Ésta considera que las distinciones de *modo* son las más importantes a la hora de diferenciar un hipónimo verbal de su hiperónimo. Esta se encuentra definida en WordNet como *una manera particular de hacer algo*, es decir, como un tipo de implicación léxica.

Vecinos. Es un conjunto conformado por vocablos que mantienen cierta relación semántica con otro en específico. Dichos vocablos son seleccionados comparando diferentes contextos locales, de tal manera que los contextos más parecidos seleccionan a los vecinos.

Bibliografía

[1] Alam, H. *et al* (2002). Extending a Broad-Coverage Parser for a General NLP Toolkit, 2002. 454-460

[2] Banerjee and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2002.

[3] Basili, Roberto; Della Rocca, Michelangelo and Pazienza, Maria Tereza (1997). Towards a bootstrapping framework for corpus semantic tagging. ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?, April 4-5, 1997, Washington, D.C., USA, 66-73.

[4] Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'01 (pp. 245{250).

[5] Briscoe, Edward J. (1991). Lexical issues in natural language processing. In Klein, Ewan H. and Veltman, Frank (Eds.) Natural Language and Speech, Proceedings of the Symposium on Natural Language and Speech, 26-27 November 1991, Brussels, Belgium, Springer-Verlag, Berlin, 39-68.

[6] Brown, Peter F.; Della Pietra, Vincent J.; deSouza, Peter V.; Lai, Jennifer C. and Mercer, Robert L. (1992). Class-based n-gram models of natural language, Computational Linguistics, 18(4), 467-479.

[7] Buitelaar, Paul (1997). A lexicon for underspecified semantic tagging. ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?, April 4-5, 1997, Washington, D.C., 25-33.

[8] Chomsky, Noam (1957). Syntactic structures, Mouton, The Hague.

[9] Choueka, Yaacov and Lusignan, Serge (1985). Disambiguation by short contexts. Computers and the Humanities, 19, 147-158.

[10] Dagan, Ido and Itai, Alon (1994). Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 20(4), 563-596.

[11] Gallant, S. (2000). Context vectors: A step toward a “grand unified representation”. In Hybrid Neural Systems, revised papers from a workshop (pp. 204{210).London, UK: Springer-Verlag.

[12] Gale, William A.; Church, Kenneth W. and Yarowsky, David (1993). A method for disambiguating word senses in a large corpus, Computers and the Humanities, 26, 415-439.

[13] Gelbukh, Alexander and Bolshakov, Igor (2001). Computational Linguistics, Models, Resources and Applications, 2001.

[14] Grishman, Ralph; MacLeod, Catherine and Meyers, Adam (1994). COMLEX syntax: Building a computational lexicon. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, 5-9 August 1994, Kyoto, Japan, 268-272.

[15] Harris, Zellig S. (1951). Methods in structural linguistics. The University of Chicago Press, Chicago, xv-384 pp.

[16] Hearst, Marti A. (1991). Noun homograph disambiguation using local context in large corpora. Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research, Oxford, United Kingdom, 1-19.

[17] Hirst, G. (1987). Semantic Interpretation and the Resolution of Ambiguity. Cambridge: Cambridge University Press.

[18] Jiang, J. and Conrath D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, Taiwan, 1997.

[19] Kaplan, Abraham (1950). An experimental study of ambiguity and context. Mimeographed, 18 pp, November 1950.

[20] Kelly Edward F. and Stone Philip J. (1975). Computer Recognition of English Word Senses, North-Holland, Amsterdam.

[21] Khellmer, G. (1991). A mint of phrases. In Aijmer, K. and Altenburg, B. (eds.) English Corpus Linguistics: Studies in Honour of Jan Svartvik, London: Longman.

[22] Kintsch, Walter and Mross, Ernest F. (1985). Context effects in word identification, Journal of Memory and Language, 24(3), 336-349.

[23] Lakoff, G., & Johnson, M.(1999). Philosophy in the esh: The embodied mind and its challenge to western thought. New York: Basic Books.

[24] Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104 (2), 211{240.

[25] Leacock, C. and Chodorow M. (1998). Combining local context and WordNet similarity for word sense identification, In C. Fellbaum editor, WordNet: An electronic lexical database, pages 265–283. MIT Press, 1998.

[26] Leacock, C., Chodorow, M. and Miller, G. (1998). Using corpus statistics and WordNet relations for sense identification, Computational Linguistics, 24(1):147-165, 1998.

[27] Leacock, Claudia; Towell, Geoffrey and Voorhees, Ellen (1993). Corpus-based statistical sense resolution. Proceedings of the ARPA Human Language Technology Worskshop, San Francisco, Morgan Kaufman.

[28] Leacock, Claudia; Towell, Geoffrey; and Voorhees, Ellen M. (1996). Towards building contextual representations of word senses using statistical models. In Boguraev, Branimir and Pustejovsky, James (Eds.), Corpus Processing for Lexical Acquisition, MIT Press, Cambridge, Massachusetts, 97-113.

[29] Lenat, Douglas B. and Guha, Ramanathan V., (1990). Building large knowledge-based systems. Addison-Wesley, Reading, Massachusetts.

[30] Lesk, Michael (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine one from an Ice Cream Cone. Proceedings of the

1986 SIGDOC Conference, Toronto, Canada, June 1986, 24-26.

[31] Lin, D. (1993). Principle-based parsing without overgeneration. In 31th Annual Meeting of the Association for Computational Linguistics (ACL 1993), 112-120, Columbus, 1993.

[32] Lin, D. (1997). Using syntactic dependency as a local context to resolve word sense ambiguity. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 64–71, Madrid, July 1997.

[33] Lin, D. (1998). Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May, 1998.

[34] Lin, D. (1998). Automatic retrieval and clustering of similar words. Proceedings of C I G C -. pp. 768-774. Montreal, Canada.

[35] Lund, K., Burgess, C., & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In Proceedings of the 17th Annual Conference of the Cognitive Science Society, CogSci'95 (pp. 660{665). Erlbaum.

[36] McCarthy, D., Koeling, R., Weeds, J. and Carroll, J. (2004). Finding predominant senses in untagged text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004.

[37] Mel'čuk, Igor A. (1987). Dependency syntax; theory and practice. State University of New York Press, Albny.

[38] Mihalcea, R. and Moldovan, D. (1999). A method for word sense disambiguation of unrestricted text. In 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), 152-158, Maryland, 1999.

[39] Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6 (1), 1{28.

[40] Miller, G., Leacock, C., Tengi, R., Bunker, R. and Miller, K. (1990). Five papers on

WordNet. Special Issue of International Journal of Lexicography, 1990, 3(4).

[41] Miller, George A.; Beckwith, Richard T.; Fellbaum, Christiane D.; Gross, Derek and Miller, Katherine J. (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3(4), 235-244.

[42] Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on lsa performance. In Proceedings of the EuroConference Recent Advances in Natura Language Processing, RANLP'01 (pp. 187{193). Tzigov Chark, Bulgaria.

[43] Pado, S., & Lapata, M. (2003). Constructing semantic space models from parsed corpora. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL'03 (pp. 128{135).

[44] Patrick, Archibald B. (1985). An exploration of abstract thesaurus instantiation. Sc. thesis, University of Kansas, Lawrence, Kansas.

[45] Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2003.

[46] Pedersen, T.; Patwardhan, S. and Michelizzi (2004).WordNet::Similarity - Measuring the Relatedness of Concepts. Appears in the Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), May 3-5, 2004, Boston, MA (Demonstration System).

[47] Picard, J. (1999). Finding content-bearing terms using term similarities. In Proceedings of the 9th Conference on European chapter of the Association for Computational Linguistics, EACL'99 (pp. 241{244). Morristown, NJ, USA: Association for Computational Linguistics.

[48] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial

Intelligence, Montreal, August 1995.

[49] Resnik, P. (1998). WordNet and class–based probabilities. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 239–263. MIT Press, 1998.

[50] Resnik, Philip (1992). WordNet and distributional analysis: a class-based approach to statistical discovery. I Workshop on Statistically-based Natural Language Processing Techniques, San Jose, California, 48-56.

[51] Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. Communications of the ACM, 8 (10), 627-633.

[52] Karlgren, J., & Sahlgren, M.(2001). From words to understanding. In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.), Foundations of real-world intelligence (pp. 294-308). CSLI Publications.

[53] Sahlgren, M. (2005). An introduction to random indexing. In H. Witschel (Ed.), Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE'05,Copenhagen, Denmark, august 16, 2005 (Vol. 87).

[54] Sahlgren, M. (2006). Towards pertinent evaluation methodologies for word-space models. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06.

[55] Sahlgren, M., & Coster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In Proceedings of the 20th International Conference on Computational Linguistics, COL-ING'04 (pp. 487{493).

[56] Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. McGraw-Hill.

[57] Salton, G., & Yang, C. (1973). On the speci_cation of term values in automatic

indexing. Documentation, 29, 351{372.

[58] Saussure, F. (1916/1983). Course in general Linguistics. Duckworth. (Translated by Roy Harris)

[59] Schütze, Hinrich (1992). Dimensions of meaning. Proceedings of Supercomputing'92. IEEE Computer Society Press, Los Alamitos, California. 787-796.

[60] Schütze, Hinrich (1993). Word space. In Hanson, Stephen J.; Cowan, Jack D. and Giles, C. Lee (Eds.) Advances in Neural Information Processing Systems 5, Morgan Kauffman, San Mateo, California, 5, 895-902.

[61] Stevenson, M. (2003). Word Sense Disambiguation: The case for Combinations of Knowledge Sources, (2003).

[62] J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. TSD 2007. 10th Internacional Conference o Text and Speech.

[63] J. Tejada-Cárcamo, A. Gelbukh, H. Calvo. Revista 40, Marzo 2008.

[64] Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In Proceedings of the 20th International Conference of Computational Linguistics, COLING'04 (pp. 1015{1021).

[65] Wilks, Yorick A. (1973). An artificial intelligence approach to machine translation. In Schank, Roger and Colby, Kenneth (Eds.). Computer Models of Thought and Language, San Francisco: W H Freeman, 114-151.

[66] Yarowsky, David (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, 23-28 August, Nantes, France, 454-460.

[67] Yarowsky, David (1993). One sense per collocation. Proceeding of ARPA Human Language Technology Workshop, Princeton, New Jersey, 266-271.

[68] Yarowsky, David (1994a). Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 88-95.

[69] Yarowsky, David (1994b). A comparison of corpus-based techniques for restoring accents in Spanish and French text. Proceedings of the 2nd Annual Workshop on Very Large Text Corpora. Las Cruces, 19-32.

[70] Zipf, G.(1949). Human behavior and the principle of least-e_ort. Cambridge, MA: Addison-Wesley.

Anexo 1

Librería WordNet::Similarity

WordNet::Similarity, es una librería que implementa medidas de proximidad semántica, las cuales se encuentran basadas en la estructura y contenido de WordNet. Las medidas implementadas en esta librería han sido divididas en medidas de relación y similitud semántica de acuerdo a los trabajos presentados por Budanitsky y Hearst.

Las medidas de similitud semántica usan la jerarquía de hiperónimos, y cuantifican cuan parecido o similar puede ser un concepto *A* con respecto a un concepto *B*. Por ejemplo, una medida de similitud debería de mostrar que *automobile* es más parecido a *boat* que a *tree*, debido a que *automobile* y *boat*, comparten a *vehicle* como antecesor en la jerarquía de sustantivos de WordNet. WordNet 2.0 tiene nueve jerarquías de sustantivos que incluyen 80,000 conceptos y 554 jerarquías de verbos que conforman 13,500 conceptos. Algunas de las medidas de similitud implementadas sólo pueden procesarse sobre vocablos que tienen la misma categoría gramatical, como los sustantivos *cat* y *dog*, o los verbos *run* y *walk*. Pese a que WordNet también incluye adjetivos y adverbios, éstos no se encuentran organizados en dicha jerarquía, de tal manera que las medidas de similitud no pueden ser aplicadas.

Sin embargo, los conceptos pueden relacionarse de muchas maneras más allá de ser similares unos con otros. Por ejemplo, *wheel* es una parte de *car*, *night* es el opuesto de *day*, *snow* es hecho de *water*, *knife* es usado para cortar *bread*, etc. WordNet proporciona otras relaciones, tales como: *parte de*, *es hecho de*, y *es un atributo de*. Toda esta información puede soportar la creación de ciertas medidas de relación semántica. Estas medidas tienden a ser más flexibles, y permiten calcular valores que cuantifiquen la relación semántica entre palabras de diferentes categorías gramaticales; por ejemplo el verbo *murder* y el sustantivo *gun*. Esta librería se encuentra disponible en la red, específicamente en *www.d.umn.edu/~tpederse/similarity.html*.

Medidas de similitud semántica

Tres de las seis medidas de similitud semántica que se implementan en esta librería están basadas en el *contenido de información* que aporta un nodo común a un par de conceptos, que en inglés se denomina *least common subsumer* (LCS). El *contenido de información* es un valor que denota la especificidad de un concepto y el LCS es un valor que toma en cuenta el nodo antecesor que proporcione mayor *contenido de información* para ambos conceptos. Estas medidas son las propuestas por Resnik (*res*), Lin (*lin*) y Jiang–Conrath (*jcn*). El corpus que por defecto fue usado para computar los valores correspondientes al *contenido de información* para cada nodo de WordNet es SemCor; sin embargo, existen programas utilitarios disponibles en la librería *WordNet::Similarity*, que permiten al usuario computar dichos valores usando el *Brown Corpus*, *Penn Treebank*, *the British National Corpus* o cualquier otro corpus.

Dos medidas de similitud están basadas en la longitud de rutas entre parejas de conceptos, específicamente las medidas propuestas por Leacock–Chodorow (*lch*) y Wu–Palmer (*wup*). La medida propuesta por Leacock–Chodorow encuentra la ruta más corta entre dos conceptos y escala ese valor por la longitud máxima encontrada en la jerarquía de hiperónimos. La medida creada por Wu–Palmer encuentra la profundidad del LCSde un par de conceptos, y luego escala dicho valor teniendo en cuenta la suma de las profundidades de cada nodo. La profundidad de un concepto es simplemente la distancia de dicho nodo al nodo raíz.

WordNet::Similarity soporta el uso de raíces hipotéticas, característica que puede ser habilitada o deshabilitada. Cuando está habilitada, un nodo raíz agrupa la información de todos los conceptos referentes a sustantivos y otro nodo raíz agrupa los conceptos referentes a los verbos. Si se encuentra deshabilitada, los conceptos de sustantivos y verbos se encontrarán en la misma jerarquía.

Medidas de relación semántica

Las medidas de relación semántica son más generales que las anteriores, ya que éstas pueden utilizarse entre palabras que tengan diferente categoría gramatical,

de tal manera que no se encuentran limitadas a usar una jerarquía específica. Tres de estas medidas han sido implementadas en *WordNet::similarity*, específicamente las propuestas por Hirst–St–Onge (*hso*), Banerjee–Pedersen (*lesk*) y Patwardhan (*vector*).

La medida propuesta por Hirst–St–Onge clasifica las relaciones en WordNet asignándoles una dirección, y luego establece una relación entre dos conceptos encontrando una ruta que no sea muy larga y que no cambie de dirección frecuentemente. Las otras dos medidas (*lesk* y *vector*) incorporan información de las glosas de WordNet.

La medida propuesta por Banerjee–Pedersen encuentra traslapes de vocablos entre las glosas de dos conceptos y además usa aquellos conceptos con los que se encuentran directamente enlazados a través de las diferentes jerarquías de WordNet. La medida propuesta por Patwardhan, crea una matriz de co-ocurrencia para cada vocablo existente en las glosas de WordNet tomando como referencia cualquier corpus de texto. Luego, representa cada glosa con un vector que es el promedio de los vectores de los vocablos co-ocurrentes.

Uso de WordNet::Similarity

La implementación de la librería *WordNet::Similarity* está basada en algunos módulos desarrollados previamente, los cuales forman parte de las librerías proporcionadas por CPAN (*Comprehensive Perl Archive Network*), tales como el paquete *Text-Similarity* usado para poder encontrar los vocablos comunes en las glosas proporcionadas por WordNet y el paquete *WordNet::QueryData* [46]**¡Error! No se encuentra el origen de la referencia.**, el cual permite crear un objeto de consulta a las bases de datos textuales de WordNet.

WordNet::Similarity ha sido implementado en Perl siguiendo los lineamientos de programación orientada a objetos. Dicha librería proporciona diversos utilitarios y métodos específicos para el uso y la implementación de nuevas medidas de similitud y relación semántica. La gráfica siguiente muestra un diagrama UML (por sus siglas en inglés *Unified Modeling Language*), en el que se especifica las clases y métodos implementados en la

librería *WordNet::Similarity*.

La utilidad *similarity.pl* permite al usuario obtener un valor que cuantifica la similitud o relación semántica entre parejas de conceptos usando una medida específica. El formato que éste utiliza para especificar un sentido específico de un vocablo ambiguo es *word#pos#sense*; por ejemplo *car#n#3* hace referencia al tercer sentido del sustantivo *car*. También permite especificar todos los sentidos asociados a un vocablo usando el formato *word#pos.* Por ejemplo en la siguiente figura, el primer comando obtienen el valor de similitud entre el segundo sentido del sustantivo *car* (*railway car*) y el primer sentido del sustantivo *bus* (*motor coach*). El segundo comando obtiene la pareja de sentidos con mayor grado de similitud para los sustantivos *car* y *bus*. En el tercer comando el argumento *allsenses* permite obtener los valores de similitud entre cada uno de los sentidos del sustantivo *car* y el primer sentido del sustantivo *bus*. En los tres comandos anteriores se usó la medida de Lin (*WordNet::Similarity::lin*).

```
> similarity.pl --type WordNet::Similarity::lin car#n#2 bus#n#1
car#n#2  bus#n#1  0.530371390319309        # railway car versus motor coach

> similarity.pl --type WordNet::Similarity::lin car#n bus#n
car#n#1  bus#n#1  0.618486790769613        # automobile versus motor coach

> similarity.pl --type WordNet::Similarity::lin --allsenses car#n bus#n#1
car#n#1  bus#n#1  0.618486790769613        # automobile versus motor coach
car#n#2  bus#n#1  0.530371390319309        # railway car versus motor coach
car#n#3  bus#n#1  0.208796988315133        # cable car versus motor coach
```

Uso de la librería WordNet::Similarity para comparar sentidos

En la siguiente figura, se crea un objeto de la clase lin y luego encuentra la similitud entre el primer sentido del sustantivo *car* (*automobile*) y el segundo sentido del sustantivo *bus* (*network bus*) usando el método *getRelatedness.*

La librería *WordNet::Similarity* proporciona la capacidad de realizar un seguimiento detallado para los procesos que computan las diferentes medidas de similitud. Por ejemplo, para las medidas basadas en longitudes de ruta, el seguimiento muestra las rutas intermedias entre los conceptos. Para las medidas que se basan en el *contenido de información* permite supervisar las rutas entre conceptos y también la búsqueda

del LCS El seguimiento para la medida *hso* muestras las rutas encontradas en WordNet, mientras que para la medida *lesk* muestra el traslape de glosas encontradas entre dos conceptos.

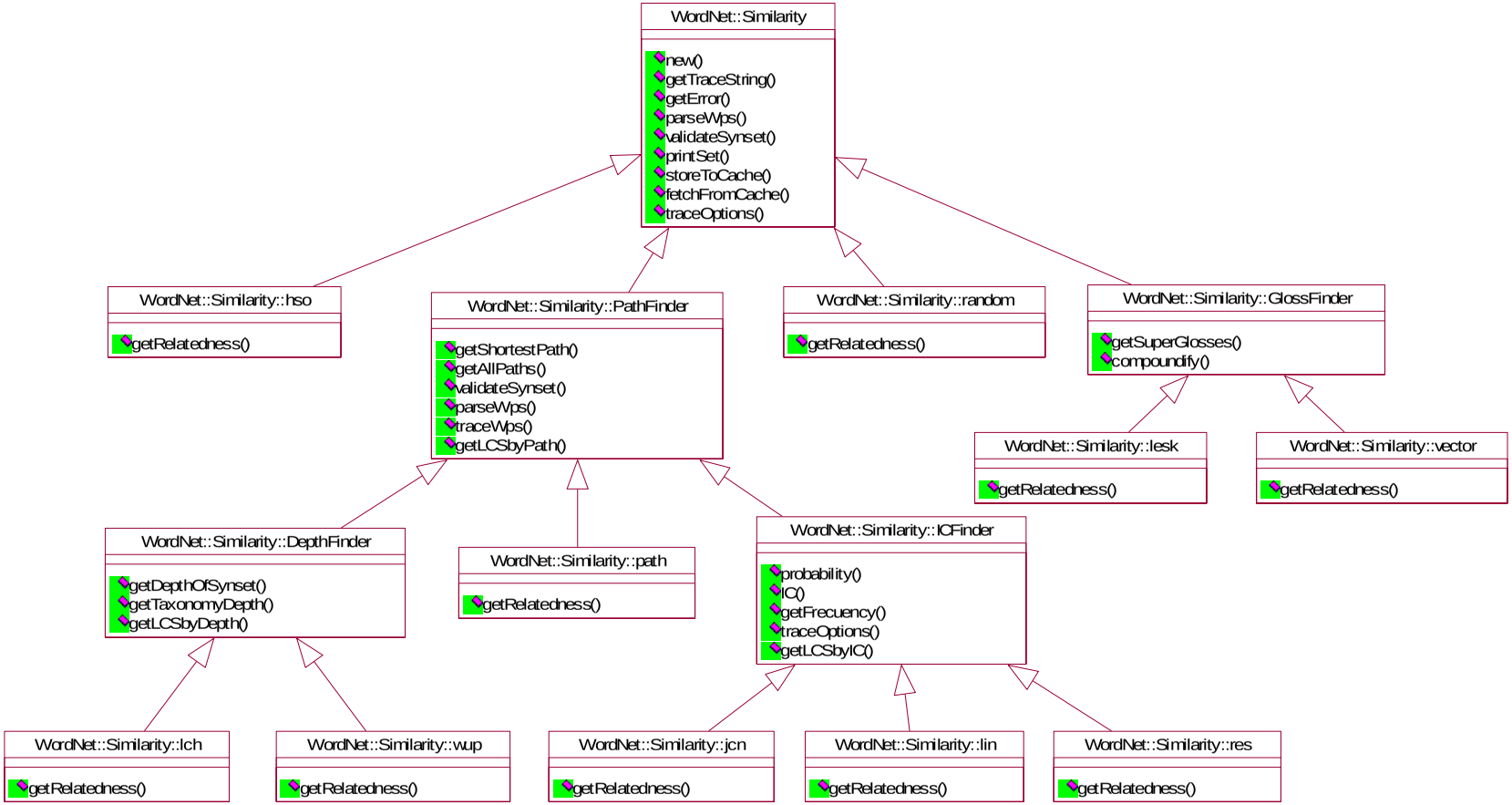
```
#!/usr/bin/perl -w

use WordNet::QueryData;                        # use interface to WordNet
use WordNet::Similarity::lin;                  # use Lin measure

 $wnObj = new WordNet::QueryData;                # create a WordNet object
$linObj = new WordNet::Similarity::lin($wnObj); # create a lin object

$value = $linObj -> getRelatedness ('car#n#1', 'bus#n#2');  # how similar?
```

Comparación de sentidos usando la librería WordNet::Similarity



. Clases implementadas en la librería WordNet::Similarity

El módulo *similarity.pm* es la superclase de todos los módulos, y proporciona servicios generales usados por todas las medidas como la validación de identificadores de *synsets*, seguimiento y almacenamiento de los resultados en la memoria de la computadora. Existen cuatro módulos que proporcionan toda la información requerida para cualquiera de las medidas soportadas: *pathfinder.pm*, *ICFinder.pm*, *depthFinder.pm* y *LCSFinder.pm*.

El módulo *pathfinder.pm* proporciona el método *getAllPaths()*, el cual encuentra todas las rutas entre dos *synsets*, y *getShortestPath()* determina la longitud de la ruta más corta entre dos conceptos en cualquiera de las jerarquías proporcionadas por WordNet.

El módulo *ICFinder.pm* proporciona el método *IC ()*, el cual obtiene el valor escalar del *contenido de información* de un *synset*. Los métodos *probability()* y *getfrecuency()* encuentran la probabilidad y la frecuencia de un *synset* basándose en cualquier corpus que haya sido usado para computar el *contenido de información*. Estos valores son calculados previamente, de tal manera que estos métodos son de sólo lectura.

El módulo *depthFinder.pm* proporciona métodos que leen valores previamente calculados por la utilidad *wnDepths.pl*, la cual obtiene la profundidad de un *synset* y su ubicación en la jerarquía de hiperónimos. La profundidad de un *synset* se calcula con el método *getDepthOfSynset()* y la máxima profundidad de una jerarquía se calcula con el método *getTaxonomyDepth()*.

El módulo *LCSFinder.pm* proporciona métodos que encuentran el LCS de dos conceptos usando tres criterios diferentes. Dichos criterios son necesarios desde que existe herencia múltiple de conceptos en WordNet y además diferentes LCS pueden ser seleccionados para una pareja de conceptos, ya que alguno de ellos puede tener múltiples padres en una jerarquía. El método *getLCSbyIC()* escoge el LCS para una pareja de conceptos que tenga el *contenido de información* más alto, *getLCSbyDepth()* selecciona el LCS que tenga la profundidad más alta y *getLCSbyPath()* selecciona el LCS que tenga la ruta más corta.