BENEMÉRITA UNIVERSIDAD
AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

UNA METODOLOGÍA PARA LA CONSTRUCCIÓN AUTOMÁTICA DE THESAURI ENRIQUECIDOS

TESIS PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA
CUPERTINO LUCERO ALVAREZ

ASESOR:
M.C. DAVID EDUARDO PINTO AVENDAÑO

PUEBLA, PUE. ENERO DE 2005

Dedicatoria

En memoria de mi padre: Anatolio Lucero Huerta, y

A mi madre: Emma Álvarez Zúñiga por todo su apoyo, comprensión y cariño.

GRACIAS

Gracias por regalarme la vida,
por traerme a este mundo de retos,
por enseñarme a vivir en paz y armonía,
sin importar sus sufrimientos.

Gracias por regalarme la vida,
por enseñarme la razón con el ejemplo,
por saber con su amor ponerme vías,
por enseñarme a terminar proyectos.

Por tanto amor, por tantas alegrías
por decirme en sus consejos
verdaderas profecías,
por enseñarme a andar subidas
para poder mirar más lejos.

QPrr.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo brindado, mediante la beca 176888.

A la Vicerrectoría de Estudios y asuntos de Postgrado (VIEP-BUAP) por el apoyo brindado, mediante la beca del proyecto: III 9-04/ING/G. Y por el apoyo para asistir a congresos de ciencias de la computación.

A la maestría en ciencias de la computación de la FCC-BUAP por el apoyo brindado, para los gastos de titulación, mediante el programa de estímulos a la titulación oportuna.

A mi asesor MC. David Eduardo Pinto Avendaño, por todo el apoyo brindado durante el desarrollo de este trabajo, y por su valiosa contribución, tanto académica como moral, durante mi estancia en la maestría.

A mis jurados: Dr. Héctor Jiménez Salazar, por su invaluable contribución, por sus acertados consejos, por su gran disposición y por su calidad humana; y Dr. Luís Villaseñor Pineda por sus contribuciones efectivas a este trabajo y por su gran disposición.

A todos mis amigos de la maestría por su amistad, y en especial a Edgar, Yatzuqui, Nahim, Ignacio, Sofía, Claudia y Paúl, por su apoyo incondicional.

A toda mi familia, por su constante apoyo moral, pero sobre todo a Dios por haberme puesto en esta familia tan maravillosa, gracias: “sin ustedes no sería nada”.

Presentación

En esta era de la información electrónica, la WEB se ha convertido en el recurso de información textual más importante, esto se debe a la gran cantidad de páginas de información de todo tipo que contiene y a la "relativa facilidad" con que esta información puede ser accedida y distribuida. Actualmente existen más de 4 billones de páginas indexadas en las máquinas de búsqueda, con información de Varios géneros, dominios y lenguajes, R. Mihalcea (2004) [47]. Día a día se incrementa el número de usuarios y la cantidad de información. Sin embargo, dicho crecimiento de información no está en relación proporcional a su disponibilidad efectiva, en otras palabras, es necesario procesarla para mejorar su aprovechamiento; de ahí la importancia de desarrollar herramientas que faciliten el acceso a la información. El Procesamiento del Lenguaje Natural (PLN) es una área de investigación que busca desarrollar sistemas de cómputo que ayuden a la solución de problemas lingüísticos y computacionales, con el objeto de aligerar grandes volúmenes de información de manera confiable, H. Jiménez-Salazar (2000) [28]. Sin embargo, la ambigüedad se encuentra en casi todas sus áreas, como son: Traducción Automática (TA), Extracción de Información (EI), y Recuperación de Información (RI) entre otras. Para tratar de resolver este problema, muchos investigadores de PLN han utilizado Recursos Léxicos (RL) tales como: *thesauri*, Bases de Datos Léxicas (BDL), y Diccionarios Electrónicos (DE), G. Rigau (2004) [22]. A este enfoque de resolución automática de la ambigüedad semántica de las palabras, que se basa en el uso de

fuentes de conocimiento externas (RL), se le conoce como "método no supervisado". A pesar de que los métodos "supervisados" logran mejores resultados en cuanto a la precisión que los "no-supervisados". Debido a la escasez de *corpora* etiquetados y a la dificultad para crearlos manualmente, muchos investigadores de PLN piensan que los primeros ya han alcanzado sus más altos rendimientos, por lo que los "no-supervisados" se están convirtiendo en el futuro de la WSD, A. Pancardo-Rodríguez et. al. (2004) [3]. Por lo tanto, debido a la importancia de los RL para el PLN, y en especial de los *thesauri* para mejorar la eficiencia de los Sistemas de Recuperación de Información (SRI), existe la motivación, para desarrollar mecanismos que los construyan de manera automática.

WordNet, EuroWordNet, y el *thesaurus* de Rogets; son RL que tienen una amplia cantidad de relaciones léxicas entre palabras. A través del tiempo, estos RL han tenido varias aplicaciones en muchos campos del procesamiento del lenguaje natural, como es el caso de la RI, WSD, y refinamiento de otros RL, entre otras, P. Rosso (2003) [46] y G. Grefenstette (1994) [20]. Sin embargo, la poca especialización de estos RL conduce a la necesidad de construirlos para dominios específicos, tarea que necesita métodos automáticos para reducir el tiempo y el esfuerzo, y permitir así su aplicabilidad a otros dominios.

Específicamente, un *thesaurus* proporciona un conjunto de términos relacionados que describen un área temática, Y. De-Castilla et. al. (2003) [50]. Aplicado a la indización de una base de datos, indica al buscador qué términos utilizar para recuperar el máximo número de documentos relevantes[1]. Los términos del *thesaurus* son utilizados por los indizadores para describir el contenido de las publicaciones con coherencia, amplitud y concisión.

Además de contener una colección de palabras relacionadas, es importante también conocer el tipo de relación semántica que existe entre ellas. sin embargo, la

[1] Los *thesauri* yudan a una mayor habilidad en la selección de elementos textuales (Documentos, contextos, sentencias o términos)

identificación automática de los tipos de relaciones semánticas es una tarea difícil de resolver, debido a la gran variedad de sentidos semánticos y a la poca cantidad de rasgos que los discriminen. Por ejemplo, los términos que se encuentran en relación de *oposición* tienen un comportamiento parecido a los que se encuentran en relación de sinonimia. De manera específica, D. Cruse (1986) [11] comenta sobre los opositivos lo siguiente:

> "... in respect to all other features, they are identical, hence their semantic closeness; along the dimension of difference, they occupy opposing poles, hence the feeling of difference."

esto también fue expresado por L. Wanner (1996) [36]: los antónimos x y y cumplen que x tenga como rasgos ABC y y tenga $AB\neg C$.

El trabajo más conocido para la identificación de relaciones semánticas, es el propuesto por M. Hearst (1992) [37], quien identifica relaciones de hiponimia mediante el uso de patrones léxico-sintácticos, en los contextos de las palabras.

En este trabajo, se propone una metodología para la identificación de relaciones semánticas entre palabras relacionadas obtenidas de un *thesaurus*, para lo cuál recurrimos al uso de patrones léxico-sintácticos, distancia de separación entre palabras y Redes de Co-ocurrencia Léxica (RCL). Las RCL fueron usadas por P. Edmonds (1997) [43] para seleccionar el sinónimo más adecuado en un contexto; aquí se usan como un mecanismo de filtrado en el proceso de determinación de antónimos.

La distribución temática de este trabajo es la siguiente:

En el primer capítulo, se presentan los conceptos generales relevantes a este trabajo, también, se citan algunas investigaciones realizadas por investigadores de PLN que tienen que ver con la detección e identificación de relaciones semánticas.

En el capítulo dos se presenta un estudio sobre las RCL con el fin de mejorar la representatividad de los términos en la tarea de identificación de relaciones semánticas.

En el capítulo tres se presenta la metodología propuesta para la identificación de relaciones semánticas a partir de pares relacionados de un *thesaurus*. Se presentan también los rasgos utilizados por dicha metodología y los resultados obtenidos en la etapa de entrenamiento.

En el capítulo cuatro se presenta la etapa de prueba de dicha metodología, y se hace un análisis de los resultados considerando la opinión de expertos en el área.

En el capítulo cinco, se presentan las conclusiones de éste trabajo.

Por último, en el capítulo seis se presentan las perspectivas, así como los primeros resultados obtenidos de la aplicación de un método para la identificación de sinónimos.

Índice general

Índice de figuras

Lista de tablas

Capítulo 1

Estado del arte en la detección de relaciones léxicas

En este capítulo se presenta, a grandes rasgos el estado actual en la detección de relaciones léxico-semánticas a partir de un *corpus*. En la primera sección se presentan los principales enfoques para la asociación automática de términos; mas adelante, en la sección 1.2 se describe el trabajo de M. Hearst para la identificación de relaciones de hiponimia; en la sección 1.3 se explica el concepto de *vecinos cercanos*; en la sección 1.4 se enuncian algunas medidas del grado de similitud o asociación de términos; en la sección 1.5 se describe el estado actual de los Recursos Léxicos; en la sección 1.6 se presentan algunos enfoques sobresalientes para la asociación de términos e identificación de relaciones léxico-semánticas; y por último, en la sección 1.7 se describe un ejemplo de la aplicación del método de G. Grefenstette para la construcción de *thesauri*.

1.1. Términos de asociación

Los *corpora*, son colecciones de documentos que comúnmente, se utilizan como recursos de información textual para el entrenamiento de herramientas de cómputo que buscan construir RL y lingüísticos que coadyuven a mejorar la eficiencia de los sistemas de Procesamiento de Lengüaje Natural.

En dichos recursos textuales, los términos de asociación pueden ser concebidos como

palabras que co-ocurren frecuentemente en los contextos, por ejemplo, en el dominio de "Sistemas Operativos", en los contextos de la palabra *memoria* aparecen de manera frecuente las palabras *disco, almacenamiento, memoria-virtual, memoria-primaria* y *memoria-secundaria*; y de manera poco frecuente las palabras *dato, bus* y *tiempo*. Las palabras más frecuentes, son llamadas *palabras asociadas de primer orden* o simplemente *asociaciones de primer orden*, G. Ruge (1991) [23]. En nuestro ejemplo, las primeras, son asociaciones de primer orden con respecto de *memoria*. Por otro lado, al observar los términos asociados de primer orden para *memoria-primaria* encontramos, entre otras, las palabras *Real* y *programa*, las cuales son *asociaciones de segundo orden* con respecto de la palabra inicial *memoria*. En ocasiones es necesario identificar asociaciones de tercero, cuarto, quinto, ó *n-ésimo* orden, como es el caso de la construcción de Redes de Co-ocurrencia Léxica, tema que será descrito en la sección 2.1.

Existe una gran variedad de enfoques para lograr la asociación automática de términos, principalmente sugeridos por investigadores de Recuperación de Información y por lingüistas, una breve caracterización de algunos de ellos se enuncia a continuación.

1.1.1. Pseudo-clasificación

Para este enfoque se requiere un conjunto de documentos de un dominio y un juego de consultas, y deben ser conocidos los documentos relevantes a cada consulta. Se utiliza un algoritmo de optimización, el cuál asigna pesos a todos los pares de términos para expandir la consulta por medio de los términos relacionados con altos pesos, de tal suerte que los resultados de la recuperación sean lo más correctos posible. Esta es la fase de entrenamiento del enfoque. Después del proceso de entrenamiento, los pares de términos con pesos altos representan asociaciones de términos. La desventaja de este enfoque recae en el alto esfuerzo, tanto para la determinación manual de juicios relevantes como para la optimización automática, G. Salton (1980) [26].

1.1.2. Asociación de términos por análisis estadístico basado en el corpus

En este enfoque se asume que la utilización de términos en un *corpus* caracteriza sus significados.

Las asociaciones de primer orden se supone que son semánticamente compatibles como "carro" y "llanta", mientras que las asociaciones de segundo orden[1] se supone que son intercambiables en el contexto y por lo tanto semánticamente similares.

En los primeros experimentos con el *corpus*, basados solamente en co-ocurrencias de términos, los documentos fueron definidos como los contextos de los términos. Ninguno de los enfoques con asociaciones de primer orden fueron usados para mejorar la recuperación por expansión automática de consultas, G. Ruge (1991) [23]. M. Lesk (1969) [39] realizó experimentos con asociaciones de primer y segundo orden, y encontró que había sólo el 20 % de términos compatibles entre las asociaciones de primer orden, pero no encontró ningún sinónimo entre las asociaciones de segundo orden. P. Lewis (1967) [44] examinó las asociaciones de segundo orden con base en contextos pequeños[2], y observó la tendencia de que los sinónimos y los antónimos se comportan más fuertes en las asociaciones de segundo orden que cualquier otro par de términos. Estos resultados[3] conducen a la conclusión de que sólo asociaciones de segundo orden en contextos pequeños pueden estar en relación semántica.

1.1.3. Asociación de términos por un análisis lingüístico basado en el corpus

Por medio de un análisis sintáctico uno puede determinar qué término está refiriendo a otro en una oración[4].

[1] Pares de términos con muchas asociaciones de primer orden en común.

[2] Usó títulos de documentos.

[3] Del uso de asociaciones estadísticas de términos.

[4] Esas relaciones de términos sintácticos constituyen en los contextos el término significativo más pequeño (un solo término).

Básicamente existen cuatro enfoques lingüísticos basados en el *corpus*, la diferencia entre ellos radica en la forma de extraer la relación sintáctica. D. Hindle (1990) [12] usaba las relaciones verbo/sujeto y verbo/objeto. G. Ruge (1991) [23] examinó asociaciones con base en la relación cabeza/modificador[5] en oraciones nominales. G. Grefenstette (1994) [20] considera ambos tipos de relaciones. T. Strzalkowski (1995) [49] examinó funciones de peso.

1.1.4. Modelo de memoria humana

Hablar del modelo de memoria humana es referirse a las investigaciones de A. Collins y E. Loftus (1975) [1], quienes presentaron uno de los primeros modelos de memoria representado como una red de activación extendida. Una red extendida es un grafo en el cuál los arcos tienen peso y los nodos están asociados con valores numéricos. Normalmente uno o más nodos son inicializados por un valor mayor a cero; estos nodos representan la descripción de un problema. La topología de la red y el peso de los arcos representan el conocimiento que es necesario para resolver el problema.

El proceso de resolver un problema de asociación, consiste en la propagación de la activación de los nodos inicialmente activados. La llamada función de activación determina la nueva activación de los nodos dependiendo de las activaciones actuales de sus vecinos y de los pesos de los arcos en común. Esta propagación es iterativa, de modo que cada iteración de una activación causa una transición de un nuevo estado en la red. El objetivo es diseñar un modelo de tal forma que un estado alcance el equilibrio después de un cierto numero de transiciones. Los nodos que son altamente activados representan la solución del problema.

[5]La relación entre dos palabras en una oración o frase, donde una refiere a la otra, es llamada relación cabeza/modificador o simplemente enlace. El modificador es la palabra que especifica a la cabeza, así pues, en la oración "dos palabras en una oración o frase", los términos: "dos", "oración" y "frase" son modificadores de la cabeza "palabra".

1.2. Patrones léxico-sintácticos

Con el objeto de apoyar a muchas tareas de PLN, el interés en la identificación automática de relaciones léxico-sintácticas y semánticas hoy día es grande. Por ejemplo, para la construcción de diccionarios, EI, WSD, generación automática de resúmenes, entre otras, conocer información de co-ocurrencia, disposición, y dependencia que existe entre las palabras en los contextos, es de mucha utilidad. Dicha información textual puede ser encontrada en los *corpora*, los cuales contienen una gran variedad de elementos informativos a cerca del lenguaje que en ellos está escrito. Por tanto, un corpus puede ser utilizado para investigar relaciones léxico-semánticas que son expresadas de maneras bien determinadas, las cuales, pueden ser identificadas de manera fácil, con muy poco entendimiento del texto.

M. Hearst (1992) [37] investiga, en *corpora* de textos grandes, relaciones de hiponimia que son directamente mencionadas en los textos, mediante el uso de patrones léxico-sintácticos como marcas y palabras clave fácilmente reconocibles, por ejemplo "Such that" o "or other".

En ese trabajo se presentan algunos patrones léxico-sintácticos y ejemplos de contextos que los cumplen, además de los pares de palabras identificadas en relación semántica de hiponimia. A continuación se presentan algunos de esos patrones:

1. *such NP as* $\{NP, \}^*\{(or|and)\}NP$

   ... works by such authors as Herrick, Goldsmith, and Shakespeare.

   $\Rrightarrow$ hyponym("author", "Herrick"),

   $\Rrightarrow$ hyponym("author", "Goldsmith"),

   $\Rrightarrow$ hyponym("author", "Shakespeare")

2. *NP* $\{, NP\}^*\{,\}$ *or other NP*

   Bruises, wounds, broken bones or other injuries ...

   $\Rrightarrow$ hyponym("bruise", "injury"),

$\Rightarrow$ hyponym("wound", "injury"),

$\Rightarrow$ hyponym("broken bone", "injury")

Los patrones léxico-sintácticos pueden ser descubiertos observando los contextos de manera cuidadosa: examinando el texto y notando los patrones y la relación que estos indican. Para encontrar nuevos patrones automáticamente, M. Hearst propone las siguientes observaciones:

1. Elegir una relación léxica de interés, por ejemplo, "grupo/miembro" palabras que están en relación de hiponimia.

2. Reunir una lista de términos para los cuales se conoce el tipo de relación de interés, por ejemplo, "Inglaterra-País".

3. Encontrar y extraer los contextos, en el *corpus*, donde estas palabras de los pares ocurren cercanas sintácticamente.

4. Encontrar los rasgos comunes a estos contextos y suponer que esos rasgos producen patrones que indican la relación de interés.

5. Una vez que un nuevo patrón se ha identificado correctamente, usarlo para reunir más instancias de la relación deseada y después volver al paso 2.

1.3. Vecinos cercanos

G. Grefenstette (1993) [19] introduce el concepto de *vecinos cercanos* para denotar a pares de términos asociados, en donde para cada par, un término es altamente frecuente en los contextos del otro y viceversa[6], por ejemplo, los contextos de las palabras **compra** y **venta** en el dominio de Economía son:

Compra $\longmapsto$ *precio 25 valor 24 venta 18 producto 15 trabajo 12 capital 12 mercancía 11 costo 10 dinero 10 beneficio 9*

[6]Un contexto de un sustantivo puede ser cualquier tipo de unidad textual (oración, sentencia, párrafo, documento, etc.), ó un conjunto de estas, en las cuales el sustantivo es parte.

Venta $\longmapsto$ *precio 42 valor 22 costo 20 compra 18 producto 16 empresa 16 mercancía 14 beneficio 13 producción 10 adquisitivo 9*

En donde se puede apreciar que en los contextos de **venta** aparece de manera frecuente la palabra **compra** y en los contextos de **compra** aparece con alta frecuencia la palabra **venta**, por lo tanto, **compra** y **venta** son *vecinos cercanos*. Los *vecinos cercanos* representan palabras más fuertemente relacionadas que las simples asociaciones de primer orden e incluso que las de segundo orden, es por eso que los *vecinos cercanos* han sido utilizados para la construcción automática de *thesauri*, este tema será explicado en el capitulo 3.

1.4. Medidas de similitud

En esta sección se describen algunas medidas de *similitud*, usadas comúnmente en Sistemas de Recuperación de Información, para discriminar entre unidades textuales, o pares de términos con algún grado de relación.

1.4.1. Medida del coseno

La manera más común para determinar la *similitud* entre dos textos (documentos sin formato)[7], representados mediante el modelo vectorial fue definida en G. Salton (1975) [25]. Los vectores representan a los documentos[8] mediante sus términos índice. Los términos índice tienen asignado un peso con base en sus frecuencias.

Sea un documento $\overrightarrow{D_i}$ considerado como un vector, esto es: $\overrightarrow{D_i} = (d_{i1}, d_{i2}, \ldots, d_{im})$, donde d_{ik} representa el peso del $k-$ésimo término en el documento i.

El peso de un término k se calcula de la siguiente manera:

$$d_{ik} = tf_{ik} \cdot (log_2(M) - log_2(df_k) + 1) \tag{1.1}$$

[7] Documentos de texto puro, sin imágenes, y sin formato.

[8] De aquí en adelante se utilizaran indistintamente los términos documento y texto.

donde, M es el número de documentos y df_k es el número de documentos que contienen el término k, es decir: es la frecuencia de un término en la colección de documentos, y tf_{ik} denota la frecuencia del términos k en el documento i.

Dados los vectores de dos documentos, $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$, se puede calcular el grado de *similitud* entre ellos: $sim(\overrightarrow{D_i}, \overrightarrow{D_j})$ calculando el coseno del ángulo:

$$sim(\overrightarrow{D_i}, \overrightarrow{D_j}) = \frac{\sum_{k=1}^{m} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \cdot \sum_{k=1}^{m} d_{jk}^2}} \tag{1.2}$$

Esta fórmula, también puede ser usada para estimar qué tan relacionadas se encuentran dos palabras en un *corpus*. Para lograr esto, es necesario representar los contextos de las palabras mediante vectores de índices, donde un ángulo de separación grande entre los vectores indica que las palabras que representan están débilmente relacionadas, mientras que un ángulo pequeño indica una relación fuerte.

Otras formas de calcular grados de *similitud* o asociación se describen a continuación:

1.4.2. Medida de Jaccard

El grado de relación entre los *vecinos cercanos* puede ser estimado mediante una variante de la medida de Jaccard[9], G. Grefenstette (1994) [20]:

$$J(a_1, a_2) = \frac{\#C(a_1, a_2)}{\#C(a_1) \cup \#C(a_1)} \tag{1.3}$$

donde $\#C(a_1, a_2)$ es el número de contextos que contienen a ambas palabras a_1 y a_2.

La medida de Jaccard calcula la proporción de que los pares co-ocurran juntos en un mismo contexto con respecto al tamaño de los contextos.

[9]Esta medida puede ser usada para estimar el grado de relación entre cualesquiera dos palabras, con base en sus contextos.

1.4.3. Puntuación t

En K. Church et. al. (1994) [35] se introduce una medida llamada "t-score" o puntuación t, la cuál es una herramienta útil para encontrar sutiles diferencias entre sinónimos cercanos. En ese trabajo se analizan pares de palabras del tipo "verbo/objeto" (V/O) con buenos resultados. El tamaño del *corpus* para sus experimentos fue de $N = 4.1$ millones con 600,000 pares diferentes.

"t-score" se define de la siguiente manera:

$$t \equiv \frac{P(V,O) - P(V)P(O)}{\sqrt{(\sigma^2(P(V,O))) + \sigma^2(P(V)P(O)))}} \tag{1.4}$$

La probabilidad de que una pareja (V/O) ocurra se puede estimar como sigue:

$$P(V,O) \approx \frac{freq(V,O) + 0.5}{N + V/2}, \tag{1.5}$$

donde $freq(V,O)$ es la frecuencia en que la pareja (V/O) aparece en el *corpus* de tamaño N, y V es el número de parejas (V/O) diferentes. Los valores: 0.5 y $V/2$ buscan un refinamiento de la fórmula clásica de probabilidad, para adaptarla a este tipo de problemas en particular.

Ahora, la varianza puede ser aproximada como sigue: $\sigma^2 P(V,O) \approx NP(V,O)$, ver: K. Church et. al. (1994) [35] para mayor información.

Por último, un par (V/O) es considerado significante, si y sólo si $t > 1.65$.

1.4.4. Información Mutua (IM)

La IM es una medida estadística muy importante en teoría de la información, por lo que ha sido muy usada en un amplio rango de aplicaciones en los últimos años, K. Church (1994) [35]. Algunas aplicaciones en RI se han discutido en C. J. Van-Rijsbergen (1999) [5].

La IM puede ayudar a distinguir los pares de palabras más interesantes de los menos interesantes, comparando la probabilidad de que concurran juntas en los mismos contextos con respecto de la probabilidad de que aparezcan en contextos distintos.

Sean x y y dos palabras, con probabilidades de ocurrencia en un texto $P(x)$ y $p(y)$ respectivamente, el grado de información mutua $I(x, y)$, entre las palabras x y y, se calcula de la siguiente manera:

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x), P(y)} \tag{1.6}$$

donde $P(x, y)$ es la probabilidad de observar x y y de manera conjunta[10], M. R. Fano (1961) [41]. Un valor alto de $I(x, y)$ indica que el grado de asociación entre x y y es fuerte.

Ahora, si consideramos como contextos de x y y a todas las oraciones del *corpus* donde estas ocurren, entonces $P(x, y) = freq(x, y)/N$, donde $freq(x, y)$ es el número de oraciones del *corpus* donde x y y co-ocurren, y N es el número total de oraciones del *corpus*[11].

Por lo tanto, la información mutua entre dos palabras x y y en un *corpus*, queda expresada de la siguiente manera:

$$IM(x, y) = \log_2 \left(\frac{N \cdot freq(x, y)}{freq(x) \cdot freq(y)} + 1 \right) \tag{1.7}$$

[10] Dos palabras ocurren de manera conjunta, si ocurren en las mismas oraciones sin importar la posición, esto es diferente de los bigramas, en los que la ocurrencia de las palabras en los contextos debe ser contigüa.

[11] Por definición clásica de probabilidad.

1.5. Recursos Léxicos

Dentro de los RL podemos englobar a los *thesauri*, los diccionarios legibles por computadora MRD (por su nombre en inglés "Machine Readable Dictionaries") y las Bases de Datos Léxicas. En muchos problemas del PLN, como son: TA, RI y EI, entre otros, el uso de RL se ha convertido en un elemento muy importante para mejorar los resultados. De esta forma, los RL se convierten en el núcleo de muchos sistemas de PLN, H. Jiménez-Salazar (2000) [28]. De modo muy general, los RL deben contener un catalogo de palabras e información sobre ellas. Esta información asociada varía de aplicación en aplicación, pero en todos los casos se trata de información lingüística útil para apoyar varios procesos en los sistemas de PLN.

En la actualidad existen algunos RL como el *thesaurus* de Rogets, el diccionarios Webster, WordNet y EuroWornet, que son muy conocidos, debido a que contienen una amplia franja de relaciones semánticas que son ampliamente aceptadas. Sin embargo, estos RL han sido creados de manera manual, lo que supone un gran esfuerzo para su creación, además de tiempo y personal calificado. Otra desventaja de esta forma de construir RL es la existencia de cierto sesgue en sus contenidos, es decir, un grupo de expertos decidió, a su juicio, qué tópicos contemplar en la jerarquía, qué palabras incluir en esos tópicos, y qué sentidos de esas palabras contemplar y cuáles ignorar. Otra desventaja es que al ser RL de propósito general, no contienen todos los ejes semánticos que uno pudiera esperar para ciertos dominios.

Los RL pueden ser utilizados para ayudar a la WSD y al etiquetamiento con partes del discurso, C. Lucero et. al. (2004) [8]. La WSD se ha desarrollado como una subárea de PLN donde el objetivo es determinar el sentido correcto de aquellas palabras que tienen más de un significado. La WSD es una tarea intermedia necesaria en varias aplicaciones de PLN y por tanto, representa en sí, el problema central a resolver en PLN, H. Jiménez-Salazar (2000) [28].

Uno de los primeros algoritmos basados en *thesauri* para desambiguar una palabra fue propuesto por D. Walker y R. Amsler (1986) [16], con base en la idea de que cada sentido y cada palabra se asignan a una o más categorías o temas en el léxico. Por lo tanto, para desambiguar una palabra, se extraían del léxico sus categorías, para posteriormente calcular la frecuencia de aparición de las palabras de la frase en las categorías y finalmente seleccionar el sentido para aquella categoría más frecuente. El léxico utilizado en esta investigación fue el *thesaurus* de Rogets. Al mismo tiempo que Walker y Amsler, M. Lesk (1986) [40] propone el uso de un MRD, para el proceso de WSD, la idea consiste en seleccionar el sentido cuya definición tenga el mayor número de empalmes con las definiciones de las palabras vecinas a la que se quiere desambiguar. Posteriormente, S. Banerjee y T. Pedersen (2002) [48] retomaron esta idea y propusieron el uso de las jerarquías de WordNet con el mismo propósito. En los últimos años han surgido otras propuestas que junto con las primeras, han convertido a los RL en una fuente de conocimiento externa de suma importancia para el proceso de WSD, específicamente, dentro de los métodos no-supervisados.

De aquí, que la necesidad de construir RL, de manera automática, ha conducido a los investigadores a desarrollar heurísticas para la asociación automática de términos e identificación de relaciones semánticas.

A continuación se describen los RL más importantes y sus características principales, más adelante se presentan las técnicas para la asociación automática de términos.

1.5.1. Thesauri

Un *thesaurus* es una colección de palabras relacionadas (términos relacionados)[12] que representan el conocimiento de un dominio[13]. Para dichos términos relacionados es muy importante conocer también el tipo de relación semántica que existe entre ellos, así como también el grado de relación, G. Ruge (1999) [24]. Normalmente los *thesauri* se utilizan en SRI para ampliar las consultas, de modo que, además de buscar por las

[12] En este trabajo, palabras relacionadas y términos relacionados se utilizan indistintamente.

[13] Área temática.

palabras que aparecen en las consultas, se buscan aquellos términos relacionados del *thesaurus*, en consecuencia se mejora la efectividad de la RI, A. Zazo et. al. (2001) [4].

En un *thesaurus*, cada palabra objetivo es introducida junto con una lista ordenada de términos relacionados. Por inspección, es posible apreciar que dichos términos relacionan a los diferentes sentidos de la palabra objetivo. Por ejemplo, en el *thesaurus* de D. Lin (D. McCarthy et. al. 2004 [14])[14] los vecinos, que se encuentran al principio de la lista, para la palabra objetivo "estrella" son: "super-estrella", "jugador", "compañero-de-equipo" y "actor", mientras que los vecinos más alejados de la lista son: "galaxia", "sol", "mundo" y "planeta" los cuales están relacionadas a otro sentido de "estrella", esto hace pensar que el grado de relación entre los términos relacionados puede ayudarnos a identificar los sentidos más comunes en el contexto.

Particularmente, el *thesaurus* de Rogets contiene una colección de más de 30,000 palabras únicas, arregladas bajo una herencia de 1000 tópicos, tales como: "existencia", "inexistencia", "sustancialidad", "insustancialidad", "canónicos" y "templos", por lo que ha sido muy utilizado en PLN: para el proceso de WSD en M. Lesk (1969) [39] y como estándar de oro de evaluación para la construcción de nuevos *thesauri* en G. Grefenstette (1993) [19].

De manera general, un *thesaurus* también puede ser usado para construir una representación normalizada de documentos para un propósito de recuperación. En este caso, los índices son solamente términos del *thesaurus* y no todas las palabras que aparecen en el documento. Para lo cuál, comúnmente se hace, de manera manual, una indexación basada en un *thesaurus*, " llamada selección de términos apropiados que caracterizan el contenido de un documento". Una desventaja de un sistema de recuperación basado en palabras que no usa una indexación basada en un *thesaurus* es que los documentos relevantes contienen otros términos tales que en la consulta no son encontrados, por tanto, la recuperación es menos eficiente.

[14]Disponible en: http://www.cs.ualberta.ca/∼lindek/demos/depsim.htm.

1.5.2. Diccionarios MRD

Los diccionarios en línea, en formato electrónico, o MRD han sido diseñados y usados por varios investigadores con el propósito de realizar descubrimientos semánticos, J. Sparck (1964) [34] y evaluar técnicas de extracción de textos no estructurados. En el trabajos de J. Sparck se intentó, por vez primera, utilizar definiciones de diccionario y sus sentidos para definir cabezas semánticas, precisamente como en el *thesaurus* de Rogets, que clasifica las palabras bajo 1,000 tópicos. Muchas palabras en diccionarios poseen un cierto número de sentidos numerados, en la investigación de J. Spark, cada sentido de diccionario fue reducido manualmente a los principales sustantivos que aparecen en la definición. La reducción fue hecha examinado frases de muestra que acompañan cada definición de diccionario, utilizando el diccionario en Inglés de Oxford, y sustituyendo todos los sustantivos que aparecen en la definición para la palabra cabeza en esa frase.

más recientemente Plate (Y. Wilks et. al. 1975) [51] realizó experimentos similares con sentidos de diccionario del Longman Dictionary, P. Proctor (1978) [45]. Este diccionario fue usado especialmente para restringir un conjunto de palabras primitivas, tales como: “niña”, “mujer”, “ceremonia”, “nación”, “relación”, “ocasión”, “clase”, etc. En las definiciones del diccionario. 2,200 primitivas fueron agrupadas usando sentidos de diccionario que co-ocurren.

En G. Grefenstette (1993) [19] se utilizó la versión 7 del diccionario Webster, como estándar de oro de evaluación[15], para evaluar pares de palabras similares. Esta evaluación se basa en el supuesto de que pares similares compartirán algunos empalmes en las definiciones de diccionario.

1.5.3. Bases de Datos Léxicas (BDL)

Las BDL son RL que, además de poseer términos relacionados, mantienen información semántica de los términos y características jerárquicas.

[15] Se utilizó como mecanismo de prueba de sus descubrimientos semánticos.

Algunas BDL como WordNet contienen un conjunto de palabras y sus definiciones, además de el tipo de relación semántica entre cada uno de los términos relacionados; otras BDL también contienen información relevante a las partes del discurso a la que cada término pertenece y algunas otras mantienen información lingüística de varios idiomas como es el caso de EuroWordNet. La importancia de la construcción de las BDL redunda en el problema central de PLN que es resolver la ambigüedad en cualquiera de sus niveles: "léxico", "sintáctico" y "semántico", H. Jiménez-Salazar (2000) [28].

WordNet fue creada en 1985 por un grupo de psicólogos y lingüistas de la Universidad de Princenton, con la finalidad de proveer a los usuarios información de diccionario ya no sólo alfabética, sino también de semántica (sinonimia, antonimia, hiponimia y meronimia). Actualmente WordNet contiene aproximadamente 95,600 formas diferentes de palabras, organizadas en alrededor de 70,100 significados o conjuntos de sinónimos (synsets) y es muy usada en PLN como es el caso de WSD, J. Morato et. al. (2004) [33]. Además de sus características propias, otras ventajas sustanciales de WordNet en relación con los *thesauri* y los MRD son: que WordNet divide el diccionario en cuatro categorías; "sustantivos", "verbos", "adjetivos" y "adverbios", y que organiza la información léxica en términos de los sentidos de las palabras, es decir, posee las características principales de los *thesauri* (agrupar las palabras bajo ciertos tópicos) y de los MRD (proveer definiciones por sentidos).

Aunque WordNet es muy usada en diferentes áreas de PLN tiene la mayor desventaja en su generalidad, ya que, aunque tiene una base de información bastante extensa, no contiene todo el conocimiento de áreas específicas.

1.6. Métodos de asociación de términos e identificación de relaciones

Como ya se ha mencionado, uno de los problemas más importantes en PLN es resolver la ambigüedad, debido a que esta se encuentra presente en casi todas sus áreas. Por lo tanto, resolver la *similitud* entre sentidos de palabras es una tarea muy importante en PLN.

Es común encontrar que la gente con frecuencia use una amplia variedad terminológica para comunicarse, incluso cuando se habla de los mismos conceptos, G. Grefenstette (1994) [20]. G. W. Furnas et. al. (1987) [27] muestra a través de un amplio rango de dominios, que la gente usa el mismo término para describir el mismo objeto con una probabilidad de menos del 20 %. No solo pueden muchas palabras diferentes ser usadas para describir el mismo concepto (ambigüedad sinonímica), sino también cada palabra individual puede tener una gran variedad de significados (ambigüedad polisémica) y lo deseable es que los sistemas de cómputo sean capaces de estimar la *similitud* e identificar cuando dos términos denotan el mismo concepto, además de seleccionar el sentido correcto de una palabra según su contexto. A pesar de que muchos investigadores han enfocado sus esfuerzos para resolver este problema, hoy día no se han logrado construir métodos automáticos que tengan un buen desempeño en términos de precisión y cobertura para WSD en dominios no restringidos, G. Rigau (2004) [22].

Los métodos de WSD básicamente, se pueden clasificar en dos tipos: supervisados y no-supervisados. Mientras que en los métodos supervisados se usan ejemplos etiquetados y algoritmos estadísticos o de aprendizaje automático para crear modelos o clasificadores de los distintos sentidos de las palabras, en los métodos no-supervisados se usan fuentes de conocimiento externas como MRD, *thesaurus* y BDL. Por otro lado, los métodos no-supervisados tienen mejores resultados en cuanto a la cobertura, los supervisados tienen mejores resultados en cuanto a la precisión, G.

Rigau (2004) [22].

Debido a que los métodos supervisados necesitan una gran cantidad de ejemplos de sentidos etiquetados para poder hacer inferencias con una precisión aceptable, y debido a que esta tarea del etiquetamiento con partes del discurso es muy compleja, tediosa y dependiente del dominio, muchos investigadores han preferido emplear métodos no-supervisados o mixtos para WSD, en los cuales se utilizan RL para consultar los diferentes sentidos de las palabras ambiguas y mediante los sentidos de sus contextos se identifica el sentido más adecuado con una eficiencia aceptable.

Los métodos supervisados de WSD, técnicas de conocimiento rico, hacen análisis morfológico de las oraciones, para lo cuál se requiere de una gran investigación de conocimiento específico de dominios completos dentro de estructuras, para después ser aplicadas a el tratamiento del texto, el costo de crear y mantener este conocimiento es grande y motiva a los investigadores a explorar otras posibles maneras de acelerar y automatizar estas adquisiciones.

Un acercamiento para la extracción automática de semántica, conocido como de conocimiento pobre, busca aprovecharse del texto que aparece en documentos cuyas estructuras semánticas son conocidas, bajo la hipótesis de que los humanos, al explicar cosas a otros humanos, entrelazan en la explicación muchos niveles de entendimiento (gramatical, sintáctico, experimental, histórico, socio-cultural, enciclopédico, etc). La tarea de las técnicas de conocimiento pobre es reconocer patrones que puedan ser explotados computacionalmente sin hacer referencia a estos niveles de entendimiento, G. Grefenstette (1994) [20].

A continuación se presenta el método de G. Grefenstette (1996) [21] para la construcción automática de *thesauri* por conocimiento pobre, y más adelante, en la sección 1.6.2, se presenta el método utilizado por D. Lin et. al. (2003) [13] para la identificación de sinónimos utilizando la WEB; en la sección 1.6.3 se verán algunos

otros métodos para la identificación automática de algunos tipos de relaciones semánticas, y por último, en la sección 1.7 se muestra un ejemplo de la aplicación del método de G. Grefenstette.

1.6.1. Método de Grefenstette

Como ya se mencionó, los *thesauri* son usados como recursos de información léxico-semántica de suma importancia para la WSD, y para mejorar el desempeño de los SRI, esto se debe a que dichos RL tiene relaciones explícitas entre palabras y relaciones implícitas entre palabras de frases sustantivadas; por ejemplo, "información" y "ciencia" son términos más generales que "ciencia de la información". Dichas relaciones pueden ser usadas para determinar el sentido correcto de una palabra en un contexto, o para hacer expanción de consultas.

Para la construcción automática de *thesauri*, G. Grefenstette propone el uso de técnicas de conocimiento pobre, específicamente, usa ventanas textuales (un trabajo para el portugués de C. Varaschin y V. Strube de Lima 2003 [10] refinó el método, incluyendo mayor información sintáctica en la representación de los términos a relacionar usando frases preposicionales). G. Grefenstette, en cuanto a la detección de relaciones semánticas, observa que usando títulos como documentos, los sinónimos nunca tienden a ocurrir juntos, pero a menudo tienden a co-ocurrir con el mismo conjunto de las otras palabras del título. Por otro lado, los adjetivos antónimos tienden a co-ocurrir en las mismas oraciones más a menudo que lo que pudiera indicar la probabilidad basada en frecuencias. Estos resultados son importantes porque indican que una cuenta simple de las palabras y las otras cadenas que ocurren con ellas pueden indicar qué palabras pertenecen a clases de significados similares. Uno de los aspectos de la variabilidad del lenguaje es que muchas palabras diferentes pueden ser usadas para describir el mismo concepto, por lo que se piensa posible encontrar un significado automáticamente de palabras asociadas con un concepto, si cada concepto es visualizado como un eje en el espacio de todos los significados, entonces una hipótesis para desarrollar este descubrimiento automático es que cuando

un concepto está siendo discutido en dos contextos diferentes habrá un amplio traslape de palabras siendo usadas para describirlo. Esta hipótesis es la base de casi todos los esquemas de comparación documento a documento y término a término en la comunidad de la recuperación de información.

El acercamiento más clásico para la extracción de semántica por conocimiento pobre usando co-ocurrencia es el que usa ventanas textuales pequeñas, por ejemplo de cuatro o cinco palabras, a la izquierda y a la derecha de la palabra objetivo. Esta técnica es fácil de implementar porque no requiere en absoluto ninguna información léxica; usualmente las palabras vacías[16] como: artículos, preposiciones, conectivos, etc. Son eliminadas[17] para reducir el tamaño de la información, los contextos extraídos de las palabras, a menudo son usados de dos maneras: para calcular qué palabras aparecen juntas con mayor frecuencia, y para observar qué palabras forman parte de los mismos contextos.

En la figura 1.1 se muestra a grandes rasgos el método de G. Grefenstette para la construcción de *thesauri*, en la cuál se pueden observar los siguientes procesos:

1. Mediante un *corpus* de dominio restringido, previamente preprocesado, se extraen los contextos de todos los sustantivos cuya frecuencia de aparición en el *corpus* exceda un umbral establecido: los contextos de los sustantivos se constituyen tomando en cuenta todas las palabras que co-ocurren con el en las mismas oraciones alrededor de una ventana establecida previamente.

2. Los contextos se forman considerando las frecuencias de las palabras en las ventanas donde los sustantivos se encuentran.

3. Se identifican los *vecinos cercanos* entre los contextos de los sustantivos: dos sustantivos son *vecinos cercanos* si uno es altamente frecuente en los contextos del otro y viceversa.

[16] Conjunto de palabras que casi no cambia en el tiempo; también conocidas como "cerradas".

[17] Al proceso de eliminación de palabras vacías y símbolos especiales (inútiles) se le conoce como "preprocesado".

4. Para todos los *vecinos cercanos*, se calcula el grado de *similitud* existente entre ellos, G. Grefenstette propone el uso de la medida de Jaccard (ec. 1.3).

5. Se calcula el tipo de relación semántica existente entre las palabras vecinas cercanas utilizando diferentes técnicas, como son: patrones léxico-sintácticos, redes de co-ocurrencia léxica, vectores conceptuales, grado de subsunción entre contextos, etc.

Finalmente se tiene un *thesaurus* que puede ser utilizado, entre otras aplicaciones, para WSD y para la expansión de consultas y contribuir a mejorar algunas tareas de PLN como la RI.

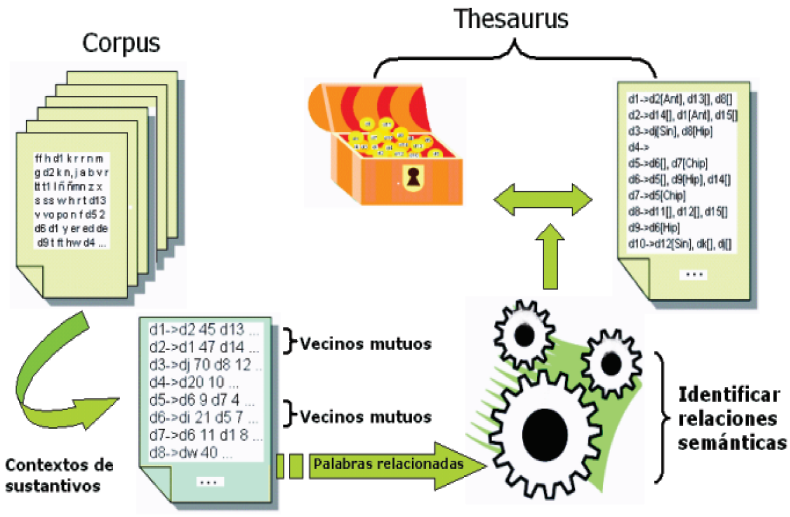

Figura 1.1: Esquema para la construcción de *thesauri* por conocimiento pobre.

1.6.2. Método de Dekang Lin

Ha habido muchas propuestas para calcular similitudes entre palabras con base en la distribución de sus contextos. Sin embargo, son pocos los enfoques que distinguen sinónimos y antónimos.

En D. Lin et. al. (2003) [13] se presenta un método para identificar relaciones semánticas de sinonimia entre palabras similarmente distribuidas. La hipótesis

distribucional indica que las palabras con significados similares tienden a aparecer en contextos similares, Z. S. Harris (1968) [53]. Considerar las palabras "adversary" y "foe". Ambas, a menudo, pueden ser usadas como los objetos de los verbos: "batter", "crush", "defeat", "demonize", "deter", "outsmart", ... y modificadores de los adjetivos: "ardent", "bitter", "formidable", "old", "tough", "worthy", ...

Dekang Lin considera los patrones de frases **from x to y** y **either x or y** para tratar de identificar la incompatibilidad entre **x** y **y**, es decir, si dos palabras **x** y **y** aparecen en uno de estos patrones, ellas muy probablemente son semánticamente incompatibles. Por ejemplo, la tabla N muestra las consultas y los logros (numero de documentos regresados) para el buscador "Alta Vista":

Query	**Hits**
adversary NEAR ally	2469
from adversary to ally	8
from ally to adversary	19
either adversary or ally	1
either ally or adversary	2
adversary NEAR opponent	2797
from adversary to opponent	0
from opponent to adversary	0
either adversary or opponent	0
either opponent or adversary	0

Tabla 1.1: Consultas y resultados.

Dada una consulta **x NEAR y**, el buscador Alta Vista regresa los documentos donde las palabras **x** y **y** aparecen cercanas una a la otra. Cuando dos palabras no están relacionadas, los logros para la consulta con **NEAR** tienden a ser bajos.

En D. Lin et. al. (2003) [13] se propone identificar pares de palabras semánticamente incompatibles investigando los patrones, descritos anteriormente, en la WEB. En ese trabajo se introduce la medida:

$$score(x, y) = \frac{hits(xNEARy)}{\sum_{pat \in P} hits(pat(x, y)) + \varepsilon} \tag{1.8}$$

Donde hits(query) es el número de aciertos regresados por Alta Vista para la consulta, P es el conjunto de patrones citados anteriormente, y ε es una constante pequeña usada para prevenir que el denominador de la fórmula sea cero (ε=0.0001). Un resultado pequeño indica que es muy poco probable que las palabras **x** y **y** sean sinónimos. Para determinar si dichas palabras con similar o distinta distribución son sinónimos, se calcula $score(x, y)$. Si el resultado es más alto que $\theta = 2,000$, el par (x, y) es clasificado como un par en relación de sinonimia.

1.6.3. Otros métodos de identificación de relaciones léxico-sintácticas

Clasificar palabras dentro de grupos semánticamente similares, es una tarea útil, pero los acercamientos estándares para clasificar los sustantivos, en términos de una jerarquía "is-a", han demostrado dificultad al ser aplicados a dominios no restringidos[18].

M. Sanderson y B. Croft (1999) [42] presentan un método para derivar automáticamente una organización jerárquica de conceptos de un conjunto de documentos. Las jerarquías son construidas usando un tipo de co-ocurrencia conocido como *subsunción*. Los términos más representativos del conjunto de documentos se clasifican, de los más generales a los más específicos, de esta manera los conceptos padre subsumen[19] a los conceptos hijos, la selección de dichos términos, en primer lugar, se obtiene de consultas capaces de recuperar los documentos, y en segundo lugar, de considerar las frecuencias de co-ocurrencia en los documentos recuperados, con base en la colección. Los términos ambiguos son considerados en entradas diferentes de la jerarquía.

[18] Las jerarquías "is-a" son difíciles de adquirir de manera manual pero favorablemente restringen los dominios.

[19] x subsume a y si los textos que contienen a y tienen una alta probabilidad de contener a x.

Otro método para extraer automáticamente jerarquías de palabras, fue el propuesto por E. Yamamoto et. al. (2004) [18]. En ese trabajo se intentan extraer jerarquías de sustantivos abstractos que co-ocurren con adjetivos en Japonés, para después compararlas con las jerarquías existentes en el diccionario electrónico EDR. Su trabajo se basa en la inclusión de relaciones de patrones de apariencia de varios *corpus*. Se concluye que la medida de similitud complementaria usada puede extraer una clase de jerarquías de los *corpora*.

En D. Hindle (1990) [12] se describe un acercamiento para clasificar palabras en Inglés de acuerdo a estructuras argumento-predicado que se observan en un *corpus* de texto plano. La idea se basa en que en Lenguaje Natural existen relaciones entre las palabras y sus construcciones gramaticales, en particular, es posible identificar qué palabras pueden ser argumentos de qué predicados[20]. Cada sustantivo puede por lo tanto estar caracterizado de acuerdo a los verbos que ocurren con el.

Se describe un método para determinar la similitud de sustantivos en la base de una métrica derivada de la distribución de sujetos, verbos y objetos en un *corpus* de texto grande, para lo cuál se utilizó un *parser* que extrae estructuras gramaticales de textos no restringidos, y se utiliza una medida de similitud, que con base en la sintaxis, muestra las estructuras semánticas notables[21].

Por otro lado, D. Schwab et. al. (2002) [15] propone la identificación de antonimia complementaria mediante el uso de definiciones de MRD y conceptos extraídos de un *thesaurus*, representando todos los posibles significados de cada concepto mediante vectores conceptuales, después usa la medida del coseno para determinar similitudes. En ese trabajo se presenta un modelo de simetría compatible con vectores conceptuales, para lo cuál, se definen funciones de antonimia que permiten la construcción de un vector de antónimos y la enumeración de sus elementos léxicos

[20] Para sustantivos, hay un conjunto restringido de verbos que aparecen como "sujeto de" o "objeto de". Por ejemplo, el "vino" puede ser "bebido", "producido" y "vendido" pero no "recortado".

[21] La medida de similitud entre sustantivos, se basa en la medida de IM entre verbos y argumentos.

potenciales. Finalmente se introduce una medida que evalúa como una cierta palabra es un antónimo aceptable de otra. La hipótesis es que los vectores codifican ideas asociadas a palabras o a expresiones. El sistema de aprendizaje automático de vectores conceptuales toma como entrada definiciones en lenguaje natural contenidas en MRD para úsus humanos. Estas definiciones son alimentadas por un analizador morfo-sintáctico que provee etiquetas y análisis de árbol[22]. Para un conjunto de conceptos elementales se construyen y comparan sus vectores conceptuales bajo la hipótesis de considerar a un conjunto de conceptos como un generador de lenguaje.

1.7. Ejemplo de la aplicación del método de Grefenstette

El método de G. Grefenstette de contexto sintáctico, fue aplicado a un *corpus* de 1.7MB de espacio en disco, el cuál está constituido de información en texto sin formato, en el dominio de "Sistemas Operativos". Después del preprocesado del *corpus*, se eligieron los sustantivos cuya frecuencia es mayor o igual que 10; posteriormente se formaron sus contextos constituidos de todas las palabras que co-ocurren con los sustantivos en las mismas oraciones y dentro de una ventana de 10 elementos: cinco a la izquierda y cinco a la derecha de cada sustantivo, debido a la posición de los sustantivos en las oraciones, si cualquiera de las ventanas no se completa, entonces se toman las palabras necesarias del otro lado, pero dentro de la misma oración, en caso de que la oración sea demasiado pequeña (por ejemplo de 5 palabras) se toman todas sus palabras para formar el contexto; una ves formados los contextos de los sustantivos, se procedió a identificar los *vecinos cercanos* dentro de una ventana de las 12 palabras más frecuentes de los contextos.

Una muestra de las palabras relacionadas obtenidas de la aplicación del método, se puede ver en la tabla 1.2.

[22]Relaciones, tales como sinonimia, antonimia e hiperonimia, que están más o menos explícitamente mencionadas en las definiciones pueden ser usadas como una manera de incrementar la coherencia de los vectores.

Sustantivo	Palabras vecinas
Sistema	operativo archivo proceso usuario tiempo programa memoria ...
Proceso	tiempo cpu usuario memoria operativo ...
Memoria	programa principal disco real sistema proceso virtual almacenamiento ...
Software	programa hardware usuario operativo ...
Hardware	programa software usuario operativo sistema memoria ...

Tabla 1.2: Palabras relacionadas en el dominio de SO.

Capítulo 2

Redes de Co-ocurrencia Léxica

Con el propósito de discriminar, de un conjunto de sinónimos, el sinónimo más adecuado en un contexto, P. Edmonds (1997) [43] propone las Redes de Co-ocurrencia Léxica como una estructura Léxico-Sintáctica, que representa las asociaciones de orden n de una palabra inicial llamada raíz. Cada nodo de la RCL tiene asignado un peso de *significación* que es inversamente proporcional a su orden, el cuál se calcula tomando en cuenta los pesos de *significación* de los nodos precedentes, desde la raíz, en su misma trayectoria. Para el cálculo de los pesos de *significación* de los nodos de las RCL, P. Edmonds propone la siguiente fórmula:

$$sig(w, x) = \frac{1}{d^3} \sum_{w_i \in P} \frac{t(w_{i-1}, w_i)}{i}, \tag{2.1}$$

donde $P = (w_0, w_1, ..., w_n)$ es considerado el camino de costo mínimo entre $w_0 = w$ y $w_n = x$, así w y x son nodos de la RCL, siendo w la raíz y x el nodo objetivo. $t(w_{i-1}, w_i)$ es la fórmula "t-score" definida en la ec. 2.7 y d es el orden de la relación.

Por ejemplo, para la RCL de la figura 2.1, el peso de *significación* para el nodo "fijo" se calcula de la siguiente manera: $Sig(costo, fijo) = \frac{1}{3^3}(t(costo, beneficio) + t(beneficio, salario)/2 +$
$+t(salario, fijo)/3) \approx \frac{1}{27}(4.57 + 1.63 + 0.44) \approx 0.246.$

Por último, para seleccionar el sinónimo más adecuado en un contexto se investigan las RCL que los representan mediante sus pesos de *significación*.

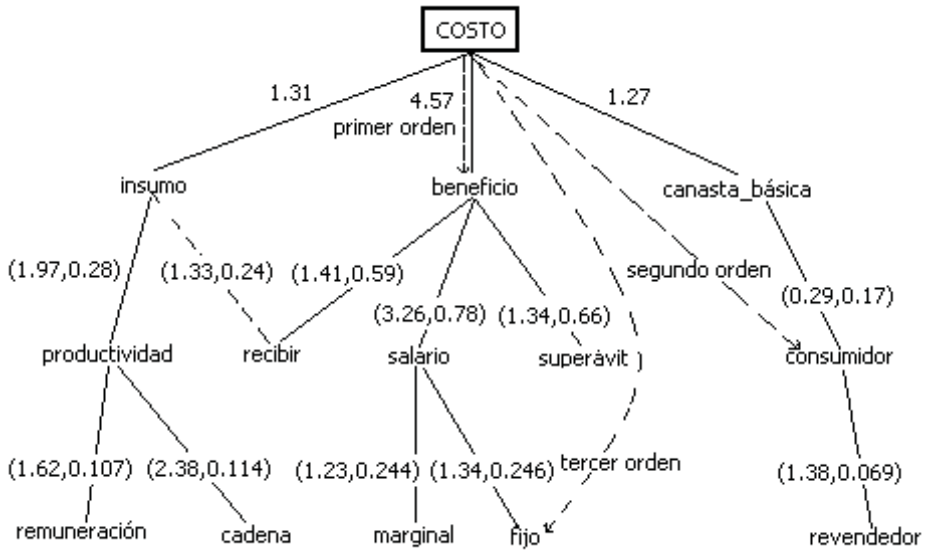

Figura 2.1: Fragmento de la RCL para la palabra *costo*

Las RCL introducidas anteriormente, pueden ser adaptadas y analizadas tomando en cuenta algunas consideraciones que apuntan a una mejor representatividad para la identificación de relaciones semánticas.

La medida de IM, ec. 1.4.4, se ha utilizado en C. Lucero et. al. (2004) [6] como un filtro de las asociaciones de primer orden en el proceso de construcción de RCL, para la identificación de antónimos. Más recientemente en C. Lucero et. al. (2004) [7] se implantó esta idea y se construyeron RCL refinadas, considerando la medida de IM dentro de un rango de valores, el propósito fue, identificar relaciones semánticas de *oposición* y de *amplitud*. En la sección 2.1 se describe el proceso de construcción de RCL refinadas, la justificación de tal refinamiento se describe en la sección 2.4.

Las RCL han demostrado ser útiles para la identificación de relaciones semánticas, pero su efectividad depende de sus características. Así, un análisis por niveles de asociación conduce a grados de *similitud* más precisos entre palabras relacionadas; mientras que usar, de manera diferenciada, las *RCL reflexivas* y las *RCL disyuntivas*,

mejora el proceso de identificación de relaciones, estos tópicos serán vistos en las secciones 2.2 y 2.3 de este capítulo.

2.1. Proceso de construcción de RCL

El proceso de construcción de RCL se describe a continuación, y más adelante, en esta sección, se enuncian las medidas fundamentales que se usan en la identificación de relaciones semánticas, lo cuál será explicado en el capítulo 3 de éste trabajo.

Sea $\mathcal{C}$ un *corpus* de dominio específico, y sea x una palabra raíz para la que se desea construir una RCL, entonces la red se construye de la siguiente manera:

1. Para formar el contexto de la raíz, x, se consideran las oraciones del *corpus* (usado en la construcción de nuestro *thesaurus*), $\mathcal{C}$, que la contienen:

$$A_1(x) = \{y | x, y \text{ co-occurren en una oración de } \mathcal{C}\} \tag{2.2}$$

2. El contexto de la raíz es filtrado, descartando todas las palabras cuya IM está fuera del rango $[4.5, 6]$, ver sección 2.4; el resto de las palabras se les llama palabras asociadas de primer-orden:

$$A'_1(x) = \{y | y \in A_1(x) \wedge 4.5 < IM(x, y) < 6\} \tag{2.3}$$

3. El proceso se repite para las palabras $y \in A'_1(x)$ (palabras asociadas de segundo-orden). Esto depende del nivel deseado de la RCL. En general, las palabras asociadas de orden n para x son determinadas de acuerdo con:

$$A'_n(x) = \bigcup_{y \in A'_{n-1}(x)} A'_1(y) \tag{2.4}$$

Una vez construidas las RCL, se calcula el peso de *significación* (ecuación 2.1) de cada nodo, para posteriormente calcular los grados de *similitud* y *contención*. Esto se logra de la siguiente manera.

Dadas las palabras a_1 y a_2 con nodos en dos RCL: $L(a_1)$ y $L(a_2)$, respectivamente, se espera que ellas tengan una alta *similitud relativa* s_r, esto ocurre si la suma de las *significaciones*[1] de las palabras en $L(a_1) \cap L(a_2)$ para ambas RCL es alta. A la suma de las *significaciones* de las palabras comunes a ambas RCL se le llama *similitud común* y es calculada tomando en cuenta sólo las palabras comunes a ambas RCL que además se encuentren en el mismo nivel de asociación: las palabras que se encuentren en ambas RCL y que no están en los mismos niveles no se consideran comunes. Por lo tanto, para saber que tan similares son dos RCL es necesario conocer la *similitud relativa*, la cuál se calcula dividiendo la *similitud común* entre la *similitud total*, donde la *similitud total* s_t, es calculada sumando todos los pesos de *significación* de ambas $L(a_1)$ y $L(a_2)$. más formalmente, la *similitud relativa* entre las palabras a_1 y a_2 está definida como:

$$s_r(a_1, a_2) = \frac{1}{s_t} \sum_{w \in \{a_1, a_2\}, x \in L(a_1) \cap L(a_2)} sig(w, x), \tag{2.5}$$

Más adelante en este capítulo se presentan algunas convenciones tomadas en cuenta, tanto en el proceso de construcción de las RCL[2], como en el proceso de análisis[3].

2.2. Análisis de RCL por niveles de asociación

Aunque en esta y las subsecuentes secciones hay referencia a ejemplos hipotéticos (fig. 2.2 y 2.3), los planteamientos que se hacen tienen sustento en RCL de palabras provenientes del *corpus* de Economía. Sea pues el siguiente razonamiento en el cálculo del grado de *similitud* entre RCL.

Para calcular el grado de *similitud* entre dos RCL, es natural pensar primero en comparar los pesos de *significación* de los nodos comunes a ambas RCL con

[1] Suma de los pesos de *significación* de los nodos de las RCL.
[2] Tipos de RCL: *reflexivas* y *disyuntivas*.
[3] Análisis de RCL por niveles de asociación.

respecto de los pesos de *significación* de todos los nodos: *similitud común* entre *similitud total*. Así, es posible estimar un grado de *similitud* normalizado, el cuál, en primera instancia parece adecuado. Sin embargo, el hecho de no contemplar, en la comparación de los nodos comunes, un análisis por niveles de asociación, puede conducir a resultados menos precisos que si esto es contemplado. Una hipótesis en los anteriores planteamientos es que dos RCL idénticas en tamaño y en disposición de sus nodos representan sinónimos más fuertes, casi perfectos[4], a diferencia de dos RCL idénticas en tamaño y contenido, pero con sus nodos dispuestos de manera diferente. Por ejemplo, para las RCL A y B de la figura 2.2, las cuales tienen 11 y 12 nodos, respectivamente, y que, excepto la raíz, todos los nodos de A están también en B, pero dispuestos en orden de asociación diferente; si se investiga el grado de *similitud* existente entre ellas sin hacer análisis por niveles, la *similitud total* es de 43 unidades, mientras que la *similitud común* es de 42, de lo que obtenemos un grado de *similitud relativa* entre ellas de 97.7 %. Es posible también estimar el grado en el que una RCL contiene a la otra, en este ejemplo, la red B contiene a la red A en un 100 %[5].

Debido a los resultados anteriores, que dicho sea de paso son bastante altos, se podría suponer que las redes A y B representan palabras en relación de sinonimia casi perfecta. Sin embargo, la distribución de los nodos de las redes no es uniforme y los puntajes de los nodos más pesados de una red, los más significantes para ella, los que deberían marcar la diferencia, son contrarestados por los puntajes de los nodos de la otra red, que son comunes pero que no se hallan en el mismo nivel, lo que conduce a aumentar los grados de *similitud* y *contención* de manera engañosa. Ahora, si calculamos nuevamente los valores, pero tomando en cuenta como nodos comunes solamente aquellos que se encuentren en los mismos niveles, podremos observar diferencias contrastantes en los resultados. Más explícitamente, mientras que la *similitud total* no cambia, la *similitud común* se ha reducido a sólo 2 unidades, por lo que ahora, el grado de *similitud relativa* es de 4.7 % y el de *contención* de 10 %. Aunque no es posible emitir un juicio acertado para estos resultados debido

[4]Los sinónimos perfectos no existen, D. Cruse (1986) [11].

[5]No se toma en cuenta la raíz.

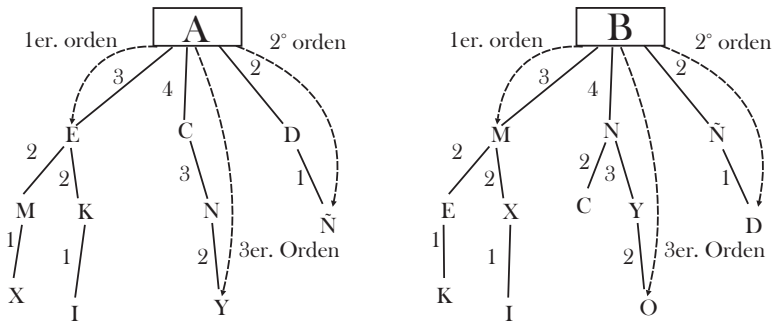

Figura 2.2: RCL tipo *disyuntivas*

a la simplicidad del ejemplo, este criterio de análisis por niveles, en la practica ha conducido moderadamente a mejores resultados.

2.3. Tipos de RCL: *reflexivas* y *disyuntivas*

En la figura 2.3 se pueden observar dos elementos gráficos, señalados mediante líneas punteadas, que son importantes en la representatividad y cálculos en las RCL, a saber: las sub-redes mútuas y la elección de enlaces. La elección de enlaces representa un aspecto muy importante en la construcción de las redes: es normal encontrar que las asociaciones de primer orden tienen muchas asociaciones de primer orden en común, si todas estas asociaciones de primer orden se tomaran en cuenta para la construcción de las redes, existirian muchas palabras comunes en los mismos niveles de asociación, por lo que la estructura de representación sería un grafo muy difícil de analizar y que no solo dificultaría los cálculos[6], sino que también no sería una buena representación del concepto central de las RCL que son las asociaciones de orden n. En otras palabras, dentro de las asociaciones de primero, segundo, tercero o n-ésimo orden es posible que, al construir las RCL, aparezcan muchos nodos iguales pero por diferentes trayectorias, por lo que, si no se realiza un proceso de elección de enlaces, es muy probable que las RCL terminen siendo complejas estructuras cíclicas. Para este trabajo se ha convenido elegir del conjunto de nodos iguales el que tenga

[6] Aumentando el tiempo de procesamiento.

el mayor peso de *significación*, respetando su trayectoria, e ignorar los de menor peso.

Por otro lado, en las RCL de la figura 2.3, en la red que representa a la raíz A se puede observar que el nodo "D" se encuentra en asociación directa de primer orden y en asociación de segundo orden a través del nodo "B", no es deseable que esto ocurra por cuestiones de sencillez y eficiencia, de sencillez por la representación, y de eficiencia porque es poco probable que el peso de *significación* de un nodo de orden superior sea más grande que el peso de *significación* de un nodo de orden inferior, esto debido a que el peso es inversamente proporcional a su orden. Ahora, en caso de que dos nodos de orden inferior tengan como asociaciones de primer orden a el mismo nodo de orden superior a ellos, se ha convenido elegir el enlace que tenga mayor peso de *significación* e ignorar el de menor peso, evitando así la proliferación de sub-redes redundantes y que se prestan a la confusión, de esta manera se garantiza que las RCL tengan la mayor *significación* posible.

Un resultado directo de las observaciones anteriores es que las RCL no tienen nodos repetidos, esto implica que el crecimiento posible de una RCL a lo más contendrá tantos nodos como tantas palabras constituyan el vocabulario del *corpus* (11575 en nuestro *corpus* de estudio). Por ejemplo, en la red que representa a la raíz A de la figura 2.3, se puede observar que los nodos "D" y "C" tienen como asociación de primer orden el nodo "M", sin embargo, el peso de *significación* del enlace (D-M) es mayor, por lo que este se establece mientras que el otro en lace (C-M) se ignora.

Una característica muy importante en los cálculos de *similitud* y *contención* entre RCL es que existen RCL que se contienen mutuamente en sus asociaciones de primer orden las cuales introducen ruido a los cálculos si sus nodos son contemplados como parte de la RCL con la que se están comparando. Dicho de otra manera, las RCL se pueden clasificar en dos grupos: *RCL reflexivas* y *RCL disyuntivas*. Dos RCL son *reflexivas* si en las asociaciones de primer orden de una se encuentra la otra y viceversa, esto es, que ambas RCL se contengan como sub-redes a partir del primer

orden. Dos RCL son *disyuntivas* si no son *reflexivas*. Estas características de las RCL han sido de suma importancia para la identificación del tipo de relación semántica entre palabras, debido a que las *RCL reflexivas* representan palabras que co-ocurren de manera frecuente en los contextos del *corpus*, mientras que las *RCL disyuntivas* representan palabras que co-ocurren con menos frecuencia juntas. Una hipótesis es que las primeras pueden ayudar a la identificación de antónimos frecuentes y las segundas a la detección de sinónimos, este tópico será analizado en la siguiente sección.

Observaciones sobre estos dos tipos de RCL como es el caso, nada sorprendente, de que más parejas en relación de *oposición* representen *RCL reflexivas* que cualquier otro tipo de relación semántica[7]; y como el hecho de que las diferencias en los tamaños de las *RCL reflexivas* sean más pequeñas que las *RCL disyuntivas*, hacen suponer que: las *RCL reflexivas* representan palabras más fuertemente relacionadas en el contexto y más estables. Sin embargo, si como subredes de otra red, no se ignoran en los cálculos de *similitud* y *contención*, el grado de *similitud relativa* se ve disminuido: esto se debe a que dichas sub-redes aportan un puntaje a la *significación* total, pero sin ningún aporte a la *significación* común. Esto es, todo el contenido de una red como subred de otra, se encuentra desfasado con respecto a ella misma. Dicho desfase privilegia la *significación* total y perjudica la *significación* común, conduciendo a resultados poco precisos. Por ejemplo, para las *RCL reflexivas* de la figura 2.3, si se toman en cuenta todos los nodos al realizar los cálculos de *similitud* (por niveles) se obtiene una *similitud total* de 68.5 unidades y una *similitud común* de 19.5, entonces el grado de *similitud relativa* entre ambas redes es del 28.46 %. Ahora, si ignoramos en los cálculos a las sub-redes mutuas tendremos una *similitud total* de 37 unidades y una *similitud común* de 19.5, lo cuál implica que el grado de *similitud relativa* entre ambas redes es del 52.7 %.

Con base en las observaciones anteriores se concluye que ignorar las sub-redes mutuas

[7]Este hecho se debe a que las palabras opuestas co-ocurren más frecuentemente en los mismos contextos que cualquier otro tipo de relación semántica.

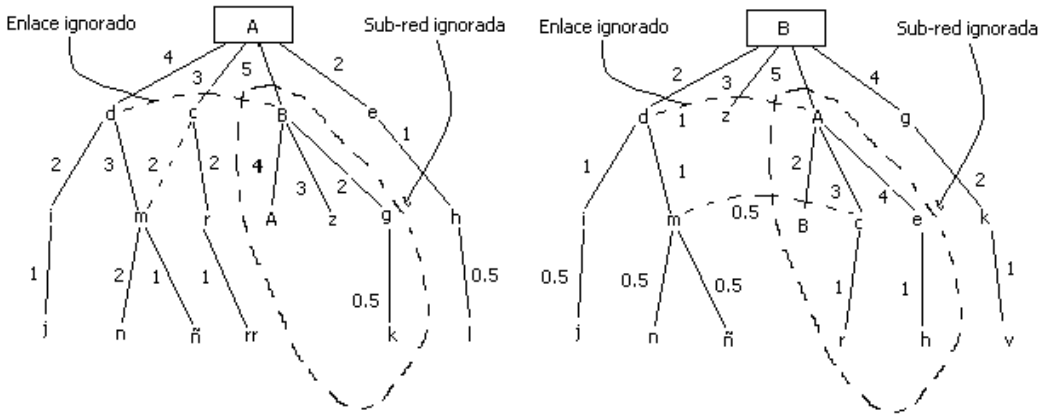

Figura 2.3: RCL tipo *reflexivas*

dentro de las *RCL reflexivas* conduce a resultados más precisos en el proceso de identificación de relaciones semánticas entre palabras relacionadas.

A continuación se presenta un experimento realizado con la medida de IM para el establecimiento del rango de valores donde ésta tiene un mayor aporte para la identificación de relaciones semánticas.

2.4. Refinamiento de RCL

La medida de IM es un indicador importante del grado de relación entre dos palabras, en este trabajo se le utiliza para filtrar los contextos de las palabras en el proceso de construcción de RCL, para lo cuál se buscó que esta medida excediera un cierto umbral. El comportamiento de las RCL construidas con diferentes umbrales de IM, se basa en experimentos realizados con el grado de *contención*[8] y se observaron los resultados de los puntajes obtenidos por las parejas que tienen alguna relación semántica de interés. Así, en el experimento se observó el comportamiento de los pares de palabras en relación semántica de *amplitud*, *oposición* y *sinonimia* dentro de las primeras 200 RCL más pesadas con respecto al grado de *contención* (medida que será analizada más adelante en este capítulo). Mientras más relaciones semánticas

[8]Proporción en que una red contiene a otra.

de esos tres tipos se encuentren entre las primeras 200, creemos que será mejor la representación de las RCL para identificar los tipos de relaciones semánticas entre los términos relacionados[9].

En la tabla 2.1 se muestran los resultados obtenidos por las RCL al variar la IM, en la primer columna se encuentran los nombres de las relaciones semánticas: *amplitud*, *oposición* y *sinonimia*, junto con "TMP" que representa palabras relacionadas y que pudieran ser términos multi-palabra que no fueron etiquetados correctamente en el *corpus*, la etiqueta "desconocida" representa a las relaciones semánticas de las cuales se desconoce su tipo de relación. En la columna $IM > 5$ se encuentran los números de palabras relacionadas de cada tipo con una $IM > 5$. Al bajar el umbral de la IM a 4.5 ya se puede observar un leve aumento en el número de relaciones interesantes y una sutil disminución del número de relaciones sin interés; al acotar la IM entre 4.5 y 7 la tendencia a mejorar se mantiene, pero los mejores resultados los obtuvimos al acotar la IM entre 4.5 y 6, obteniendo 39 pares de palabras en relación de *amplitud* de un total de 94, 17 en relación de *oposición* de un total de 47, y 9 en relación de *sinonimia* de un total de 19, mientras que el número de pares de palabras sin interés disminuyó paulatinamente. Por tanto, podemos concluir que los pares de palabras cuya medida de información mutua está dentro del rango $4.5 < IM < 6$ tienen mayor oportunidad de tener algún tipo de relación semántica de interés[10], por lo que este rango es usado para seleccionar las asociaciones de primer orden en el proceso de construcción de RCL.

Esta ha sido una prueba muy sencilla pero que nos ha conducido a construir RCL más representativas.

[9] En *thesauri* más grandes es probable que dicho número de RCL análizadas deba ser mayor.
[10] En *corpora* más grandes, talvez este rango no se mantenga.

Relación	$IM > 5$	$IM > 4.5$	$4.5 < IM < 7$	$4.5 < IM < 6$	**valor real**
Amplitud	17	19	26	39	94
Oposición	10	10	11	17	48
Sinonimia	3	4	9	9	19
Total: de interés	**30**	**33**	**36**	**65**	**161**
TMP	38	36	34	33	123
Desconocida	132	131	120	102	351
Total: sin interés	**170**	**167**	**154**	**135**	**474**

Tabla 2.1: Experimentos con la medida de IM.

2.5. Diferencias de tamaños en RCL y grado de contención

En H. Jiménez-Salazar (2003) [29] se presenta un método para la identificación automática de relaciones léxicas mediante la noción de subsunción, M. Sanderson y B. Croft (1999) [42]. La investigación compara diferentes combinaciones de contextos obtenidos mediante sus diferentes sentidos, para lo cuál se calcula el grado en que los contextos de un sentido contienen a otro, tomando en cuenta sus tamaños. En ese trabajo, se proponen algunas reglas para la identificación de relaciones semánticas con base en el grado de subsunción[11]. Dichas reglas, trasladadas al uso de RCL en lugar de los contextos de los diferentes sentidos de las palabras, quedan de la siguiente manera. Dadas dos palabras A y B,

1. Son sinónimos si el grado de *contención* de sus redes es grande y la diferencia en los tamaños de las redes es pequeña.

2. A es hipónimo B, si el grado de *contención* de la red de A en B es grande y la diferencia en los tamaños de las redes es grande.

3. Están fuertemente relacionadas si la *contención* de sus redes es grande y la diferencia en los tamaños de las redes es mediana.

[11] x subsume a y si los textos que contienen a y tienen una alta probabilidad de contener a x.

4. Están débilmente relacionadas si la *contención* de sus redes es pequeña y la diferencia en los tamaños de las redes es mediana.

5. Están muy débilmente relacionadas si la *contención* de sus redes es pequeña y la diferencia en los tamaños de las redes es grande.

En éste trabajo, creemos que el inciso a), también pudiera cumplirse para las relaciones de antonimia y co-hiponimia, mientras que el inciso c), pudiera indicar hiponimia.

Tomando en cuenta las reglas anteriores y las nuevas inferencias, las diferencias de tamaño y el grado de *contención* entre RCL, serán usados en la metodología para la identificación de relaciones de *amplitud* en las etapas de entrenamiento y prueba.

Capítulo 3

Etapa de entrenamiento

En este capitulo se presenta un nuevo método para la construcción automática de *thesauri* enriquecidos con algunos tipos de relaciones semánticas, el cuál se basa en los resultados obtenidos por el método de G. Grefenstette para la asociación automática de términos (ver sección 1.6.1), y en el uso de rasgos léxico-sintácticos para la identificación de algunos tipos de relaciones semánticas entre dichos términos. Se presentan también los resultados obtenidos de la aplicación del método al dominio de Economía. Como veremos, en el capítulo siguiente, el ajuste de parámetros hecho en este dominio es la base para extrapolar el método a otros dominios. Por tal motivo llamamos al procedimiento descrito en el presente capítulo, etapa de entrenamiento[1].

En la figura 3.1 se muestra, de forma panorámica, la arquitectura usada para la construcción del *thesaurus*. En el esquema se pueden apreciar: los recursos textuales necesarios (un conjunto de sustantivos y un conjunto de documentos o *corpus*), y los procesos principales que se llevan a cabo para la generación del recurso. más explícitamente, a partir de un *corpus* se identifican los sustantivos cuyas frecuencias excedan un umbral previamente establecido, después se identifican los *vecinos cercanos*, y dos tipos de relaciones semánticas entre ellos (de *oposición* y de *amplitud*).

Por lo tanto, el método propuesto, para la construcción automática del *thesaurus* se divide en dos etapas principales: identificación de las palabras relacionadas más

[1] Los umbrales que se establecen en este capítulo son usados en la etapa de prueba.

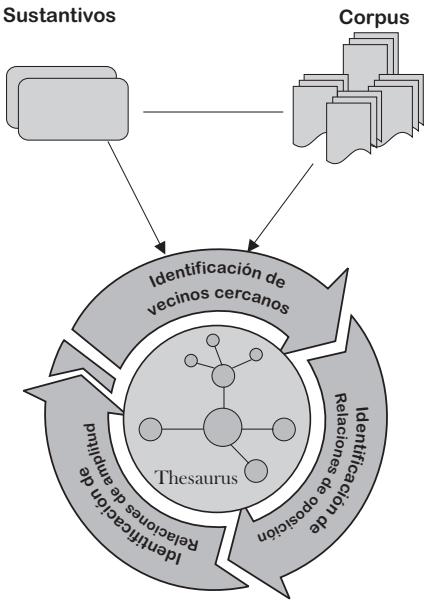

Figura 3.1: Arquitectura para la construcción automática de *thesauri*

descriptivas del dominio, e identificación del tipo de relación semántica existente entre ellas.

En la sección 3.1 se describe la forma en la que se construyó el conjunto de términos relacionados, y en la sección 3.2 se presenta la metodología empleada para la identificación automática de los tipos de relaciones semánticas. Por último, en la sección 3.3 se presenta, como resultado de la metodología, el *thesaurus* construido en el dominio de Economía.

3.1. Identificación de pares relacionados

La identificación de las palabras relacionadas fue hecha mediante análisis estadístico basado en el *corpus*, específicamente, se usó la técnica de ventanas de G. Grefenstette. Cada par relacionado se obtiene bajo la idea de *vecinos cercanos* (ver sección 1.3). El grado de relación entre las palabras relacionadas fue estimado mediante una variante de la medida de Jaccard (sección 1.3).

Se usó una lista de 32,000 sustantivos de dominio general y la regla: $Pr((v_{i-1}, m_i, v_{i+1}) \in \{el, al\} \times \{NOM\} : \{de, del\} \mid (v_{i-1}, v_{i+1}) \in \{el, al\} \times \{de, del\}) = 1$, para identificar sustantivos de propósito específico, H. Jiménez-Salazar (2000) [28]. Así, se conformó una lista de cerca de 33,000 sustantivos que fueron usados para identificar las palabras relacionadas.

Para este experimento se usó un *corpus* de 4.67MB de texto sin formato, el cuál fue truncado, utilizando el algoritmo de Porter[2], y preprocesado para eliminar palabras cerradas[3] y signos de puntuación. El *corpus* resultante, después del preprocesado, se compone de 26,297 oraciones y 11,575 términos.

Para la identificación de los *vecinos cercanos*, se formaron los contextos de los sustantivos cuya frecuencia[4] fue mayor o igual que 10, y se utilizó una ventana de entre las 14 palabras más frecuentes. El contexto de un sustantivo fue constituido tomando en cuenta el conjunto de todas las palabras que aparecen en las oraciones del *corpus*, y que co-ocurren con él. Todos los contextos son arreglados linealmente, de tal forma que todas sus palabras aparezcan listadas de manera decreciente con respecto a sus frecuencias. Una vez identificadas las palabras relacionadas se calcula el grado de relación entre ellas utilizando la medida de Jaccard (ver sección 1.3).

En la figura 3.2 se presenta la arquitectura general para la construcción de la colección de *vecinos cercanos*, para lo cuál, se necesita de un corpus preprocesado y un conjunto de sustantivos en ese dominio, esta información es usada por los procesos que construyen y ordenan los contextos, así como también, por los procesos que identifican los *vecinos cercanos* y calculan el grado de relación entre ellos.

Por último, en la tabla 3.1 se presenta una muestra de los 635 pares de palabras relacionadas obtenidas, el grado de relación y el tipo de relación semántica existente

[2]Disponible en: Http://www.tartarus.org/~martin/PorterStemmer/.

[3]Preposiciones, conectivos, verbos, adverbios, etc.

[4]Este filtro permite descartar palabras poco frecuentes y aligerar el procesamiento.

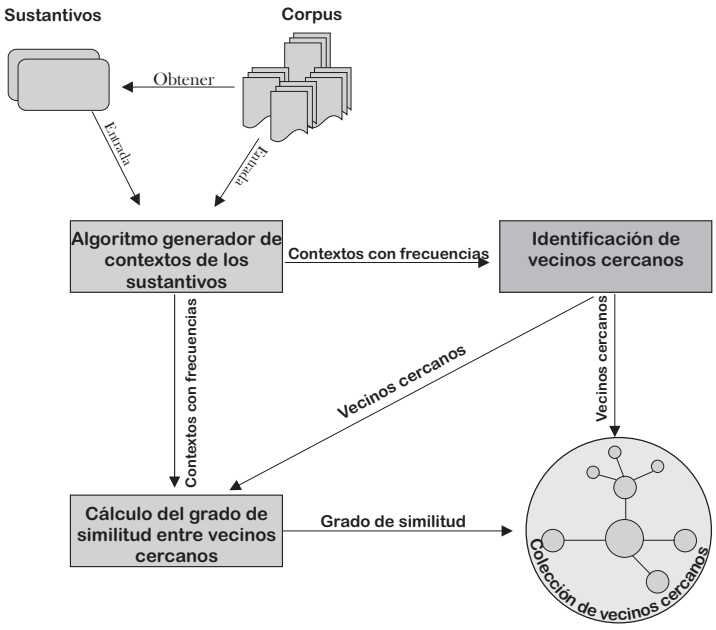

Figura 3.2: Arquitectura para la identificación automática de *vecinos cercanos*

entre ellas. Es importante señalar que el tipo de relación fue previamente etiquetado manualmente para entrenar los algoritmos de identificación automática, donde la etiqueta ND indica que para ese par relacionado se desconoce el tipo de relación, mientras que las etiquetas: "Hiper", "Holo", "Hip", "Chip", "Ant" y "Sin", indican relaciones de "Hiperonimia", "Holonimia", "Hiponimia", "Co-hiponimia", "Antonimia" y "Sinonimia" respectivamente, por ejemplo, el par ("Plata", "Metales-preciosos") indica que "Plata" es hipónimo de "Metales-preciosos". Por otro lado, el grado de relación máximo posible es igual a 0.5, de esta manera, por ejemplo, el grado de relación normalizado, entre las palabras "Plata" y "Metal" es de aproximadamente 51 %, mientras que entre las palabras "Metal" y "Moneda" es de 26.2 %.

A continuación se presenta una metodología para la identificación de relaciones semánticas entre los pares de términos relacionados, los cuales se han identificado previamente en ésta sección.

Palabra A	Palabra B	Peso	Relación
Metal	**Plata**	0.255	Hiper
Metal	**Moneda**	0.131	Hiper
Oro	**Valor**	0.189	ND
Oro	**Mina**	0.158	ND
Oro	**Moneda**	0.154	Holo
Plata	**Metales-preciosos**	0.187	Hip
Plata	**Cobre**	0.168	Chip
Plata	**Pieza**	0.149	ND
Pescado	**Aceite**	0.151	Chip
Harina	**Trigo**	0.145	Holo
Positivo	**Negativo**	0.222	Ant
Positivo	**Normativa**	0.177	ND
Externo	**Interno**	0.244	Ant
Interno	**Mercado**	0.099	ND
Explícito	**Implícito**	0.182	Ant
Justicia	**Equidad**	0.135	Sin
India	**China**	0.178	Chip

Tabla 3.1: Palabras relacionadas obtenidas del *corpus* de Economía .

3.2. Metodología para identificar relaciones semánticas

Identificar de manera automática relaciones semánticas entre palabras es la tarea más difícil en la construcción de un *thesaurus*, esto se debe, entre otras cosas, a la gran diversidad de maneras que existen en nuestra lengua para expresar una idea y a la gran diversidad de conceptos que pueden ser expresados por una sola palabra, es decir: a la ambigüedad sinonímica y a la ambigüedad polisémica, E. F. Carcedo (2003) [17], sin embargo, el hecho de que algunas palabras co-ocurran de manera frecuente en los mismos contextos hace pensar en una proximidad entre sus sentidos.

Debido a lo anterior, para la identificación del tipo de relación semántica entre las palabras relacionadas, sólo se han contemplado relaciones semánticas de *oposición* y de *amplitud*. Como relaciones de *oposición* se han considerado los antónimos complementarios (como "verdadero" y "falso"), relaciones de *oposición* (como de dirección, reversivos, antipodales, entre otros, D. Cruse (1986) [11]). En las

relaciones de *amplitud* se han incluido a las jerárquicas: "hiponimia", "hiperonimia", "holonimia", "meronimia", "co-holonimia"y "co-hiponimia". Un trabajo previo a este se describe en C. Lucero et. al. (2004) [9].

La metodología se divide en dos etapas principales: identificación de relaciones de *oposición* e identificación de relaciones de *amplitud*. En ambos procesos de identificación se usan los rasgos: RCL, DPS y PLS. La diferencia del empleo de estos rasgos radica en que en la etapa de identificación de relaciones de *oposición*, se utilizan como puntuación parcial de una función de puntuación global; y en la etapa de identificación de relaciones de *amplitud*, se utilizan como estrategia de agrupamiento y eliminación.

A continuación, en la sección 3.2.1, se describen los rasgos, y en las secciones 3.2.2 y 3.2.3 se describen las etapas de identificación de relaciones. Por último, en la sección 3.3, se describe un mecanismo para la discriminación de los resultados obtenidos en ambas etapas, además se presentan los resultados finales de la metodología para la etapa de entrenamiento.

3.2.1. Descripción de los rasgos utilizados por la metodología

Los rasgos usados por la metodología y sus aspectos más sobresalientes, se describen a continuación.

Rasgo: Redes de Co-ocurrencia Léxica

Las RCL fueron usadas en este trabajo de dos formas: 1) como una medida que indica qué tan similares son las palabras que las representan, bajo la hipótesis de que dos redes altamente similares podrían señalar una posible relación de sinonimia o antonimia; y 2) como una medida del grado de *contención* entre ellas, la cuál podría señalar alguna relación de jerarquía.

En este trabajo se ha construido una RCL para cada palabra relacionada del

thesaurus, tomando en cuenta las consideraciones vistas en el capítulo 2. El nivel de profundidad utilizado fue de tres, es decir, las RCL sólo contienen asociaciones de primero, segundo y tercer orden. Algunas pruebas realizadas en RCL con ordenes superiores a tres, mostraron que el aporte a los grados de *similitud* es mínimo con respecto a las RCL de tercer orden, además de que se incrementa sustancialmente el tiempo de procesamiento[5].

A continuación se presenta otro rasgo utilizado por la metodología para la identificación de relaciones semánticas.

Rasgo: Patrones léxico-sintácticos

Los PLS han sido introducidos en el trabajo de P. Hearst (1992) [37] con buenos resultados en la identificación de hipónimos. En este trabajo, se identificaron diversos patrones en los contextos con relaciones de *oposición* y de *amplitud*, guiados por palabras clave, signos de puntuación, y distancia entre palabras relacionadas. Los PLS son representados como expresiones regulares. Ejemplos de patrones para relaciones de *oposición* y de *amplitud* aparecen en las tablas 3.2 y 3.3[6], respectivamente. Los pesos fueron asignados con base en la precisión de cada expresión regular. El criterio consistió en asignar un punto a la ER por cada 5 puntos porcentuales de su precisión; de esta forma, una ER con peso iguál a 20 significa que esa ER tuvo una precisión del 100 %.

Nr	Expresión Regular	Peso
1	Ant word*, **pero** word* Ant	16
2	**desde** word* Ant **hasta** word* Ant	12
3	Ant word* [,|;] **sino** word* Ant	10
4	Ant word{0,4}[**y**|**o**] word{0,4} Ant	2

Tabla 3.2: ER para relaciones de *oposición* y sus pesos.

[5]En *corpora* más grandes, es probable que el orden de asociación requerido sea superior

[6]Las ER completas se encuentran en las tablas D.1 y E.1 de este trabajo (Apéndice).

Nr	Expresión Regular	Peso
1	Hip word{0,1} **a semejanza de** word{0,5} Hip	20
2	Hip, **incluyendo** word,* [o | y] Hip	12
3	**tal** Hip **como** {word,}* [o | y]{0,1} Hip	20
4	Hip word{0,8}, word{0,4}**[con el nombre de | ...]** word{0,2} Hip	20
5	Hip [word | ,]{0,18}, **[o otro |u otro |y otro]**word{0,5} Hip	12

Tabla 3.3: ER para relaciones de *amplitud* y sus pesos.

Rasgo: Distancia Promedio de Separación

Este rasgo tiene su base en la observación de contextos que tienen palabras relacionadas. DPS representa qué tan cercanas están dos palabras, es decir: es el número promedio de palabras que se encuentran entre otras dos palabras en la misma oración. En contextos de palabras relacionadas, las palabras opuestas muy frecuentemente co-ocurren con una distancia más pequeña que las que están en relación de *amplitud*. Esta observación está basada en el uso de antónimos con propósitos de contraste.

En experimentos realizados con el fin de determinar las DPS entre palabras de la misma clase, se analizaron los contextos de los 635 pares de palabras relacionadas del *thesaurus* de Economía, para lo cuál se agruparon las parejas con relación de: *oposición*, *sinonimia*, *amplitud*, *términos multi-palabra* (TMP) y parejas en relación semántica *desconocida*.

En la tabla 3.4 se muestran las clase de pares de palabras analizadas, la DPS, el número de pares, el rango de valores de las DPS donde cae el 80 % de los pares analizados, y el número de pares cuya DPS cumple con el rango. En la fila inferior de la tabla, se puede observar el promedio $\bar{x}$ de las DPS y su desviación estándar S, además del total de pares de la muestra, y el total de pares que cumplen con los rangos.

Clase	DPS	Pares	Rango	Pares en rango
oposición	2.4	47	[1,4]	38
amplitud	3.7	94	[1,6.5]	75
sinonimia	3.8	20	[1.9,5.4]	16
muti-palabra	2.5	123	[0.3,5.9]	98
desconocida	3.8	351	[1.06,6.14]	281
	$\overline{x}$ =3.43 S = 2.04	**Total = 635**		**Total = 508**

Tabla 3.4: **DPS** y rangos de valores.

Es importante señalar que para encontrar estos rangos se descartaron las DPS con valores pequeños (10 % del total), y las DPS con valores grandes (restante 10 %) del rango (que se está calculando), de esta manera, se obtuvieron los rangos donde cae el 80 % de las DPS.

Los resultados de este experimento, validan la hipótesis de que los pares de palabras en relación semántica de *oposición* co-ocurren con menor distancia en los contextos que cualquier otro tipo de relación. en algunos casos de la metodología, los valores de los rangos y las DPS, según la clase a la que corresponden, son usados como umbrales.

Determinación de rangos para otros dominios

Para determinar los rangos de valores, donde cae la mayoría de los pares según su clase, es necesario establecer los factores F_{LI} y F_{LS} para los límites inferiores y superiores respectivamente; tomando en cuenta los rangos, el promedio y la desviación estándar, observados en la tabla 3.4.

Por ejemplo, F_{LI} y F_{LS}, para determinar el rango de relaciones semánticas de *oposición*, se calculan de la siguiente manera:

$$F_{LI} = \frac{\overline{x} - LI}{S} = 1.19, \tag{3.1}$$

y

$$F_{LS} = \frac{LS - \overline{x}}{S} = 0.27, \tag{3.2}$$

mientras que los factores para establecer el rango de relaciones de *amplitud* son:

$$F_{LS} = \frac{LS - \overline{x}}{S} = 1.5, \tag{3.3}$$

y

$$F_{LI} = \frac{\overline{x} - LI}{S} = 1.19 \tag{3.4}$$

Bajo la hipótesis de que estas proporciones se mantengan al variar el *corpus*, para otro dominio, *Dom*, podemos establecer el rango de umbrales como:

$$R = [\overline{x}_{Dom} - S_{Dom} \cdot F_{LI}, \overline{x}_{Dom} + S_{Dom} \cdot F_{LS}], \tag{3.5}$$

donde $\overline{x}_{Dom}$ y S_{Dom} son: la DPS de todos los pares relacionados de la colección (*thesaurus* T_S) y la desviación estándar de dichas DPS para el dominio *Dom*, respectivamente.

3.2.2. Identificación de relaciones de *oposición*

Para la identificación de relaciones de *oposición*, se hace uso de los rasgos previamente descritos en una función de puntaje global, y de umbrales establecidos mediante experimentación. Los PLS mostraron ser el rasgo más importante, por lo cuál se utiliza primero, para seleccionar el grupo de pares de términos relacionados del *thesaurus* con mayor peso y, después, como una puntuación parcial que contribuye a la puntuación global. En este trabajo se descubrieron 22 PLS para detectar relaciones de *oposición*, mediante los cuales se seleccionó un grupo de 65 pares[7], los cuales sobrepasaron el umbral establecido[8] de 1.4

[7] Candidatos fuertes a estar en relación de *oposición*.

[8] Este umbral fue establecido por inspección, buscando acotar el mayor número de palabras opuestas dentro de un grupo lo más reducido posible de parejas, privilegiando la precisión sin descuidar mucho la cobertura.

Determinación del umbral de selección para otros dominios

Sea $U = 1.4$ el umbral de selección de pares candidatos, observado en la etapa de entrenamiento. Para calcular el umbral de selección U' para un nuevo dominio, es necesario establecer un factor f en términos de U y del promedio de los puntajes aportados por las expresiones regulares $\frac{\sum P_{ER}}{n}$, con base en el número total de pares relacionados n observados en la etapa de entrenamiento:

$$f = \frac{n \cdot U}{\sum P_{ER}}, \tag{3.6}$$

donde $n = 633$ y $\frac{\sum P_{ER}}{n} = 0.48$, por lo tanto, $f = 2.92$. Ahora:

$$U' = \frac{\sum P'_{ER}}{n'} \cdot f, \tag{3.7}$$

es el umbral de selección para la identificación de relaciones de *oposición* en otro dominio, y $\frac{\sum P'_{ER}}{n'}$ es el promedio del puntaje aportado por las ER con base en el número total de pares relacionados del thesaurus del nuevo dominio.

Cálculo del puntaje aportado por PLS

Sean Ref y Dis las clases de pares de palabras relacionadas del *thesaurus* T_S, según su tipo de RCL; *RCL reflexivas* y *RCL disyuntivas*, respectivamente. Sea $G_R \in \{Ref, Dis\}$ cualquiera de esas dos clases, con $(a, b) \in G_R$ un par de ese grupo, entonces el cálculo del puntaje relativo P_r de cada pareja de G_R se realiza de la siguiente forma:

$$P_r(a, b) = \frac{1}{N_{(a,b)}} \sum_{e \in E} weight(e) \cdot fr(e) \tag{3.8}$$

donde E es el conjunto total de ER; $weight(e)$ es el peso asignado a la expresión regular e, ver tabla D.1 (del apéndice); $fr(e)$ es el número de veces en los cuales e empata dentro de los contextos de (a, b), y $N_{(a,b)}$ es el número total de contextos donde a y b co-ocurren. La división por $N_{(a,b)}$ garantiza que P_r se encuentre

normalizado con base en el número de contextos comunes; con esto, los pares menos comunes no se ven afectados.

El puntaje relativo P_r se usa para calcular el puntaje total W_{PLS}, para lo cuál se procede de la siguiente manera:

1. Cálculo del promedio $\overline{x}$ de todos los $P_r(a,b)$.

$$\overline{x} = \frac{\sum\limits_{(a,b)\in G_R} P_r(a,b)}{\#G_R} \tag{3.9}$$

   donde, $\#G_R$ representa el número total de pares relacionados del grupo G_R.

2. Cálculo de la desviación estándar S con respecto a la media, de todos los pares de G_R:

$$S = \sqrt{\frac{\sum\limits_{x_i\in P_r(a,b)} (x_i - \overline{x})^2}{\#G_R - 1}} \tag{3.10}$$

3. El umbral de proporción $U_p = \overline{x} + S$ se hace corresponder a uno, entonces el puntaje aportado por PLS (a cada par) se calcula de la siguiente manera:

$$W_{PLS}(a,b) = \frac{Pr(a,b)}{U_p} \tag{3.11}$$

4. El puntaje aportado por los pares cuyo $P_r > U_p$ se hace corresponder a uno, de esta manera se garantiza que el puntaje parcial aportado por PLS esté normalizado.

Cálculo del puntaje aportado por DPS

La distancia promedio se utiliza en las *RCL reflexivas* y en las *RCL disyuntivas* como un puntaje determinado al complementarla con respecto al valor 4, que es el límite superior del rango de valores donde caen los promedios de la mayoría de las palabras en relación de *oposición*. La distancia complementada arroja un puntaje menor o

igual que uno, de manera proporcional. La proporción se calcula con respecto al máximo Δ_M de la distancia complementada:

$$\Delta_M = \max_{(a,b)\in G_R} \{4 - \bar{\Delta}(a,b)\}, \tag{3.12}$$

donde $\bar{\Delta}(a,b)$ es la distancia promedio entre las palabras a y b en sus contextos, y G_R representa los pares de palabras según el grupo que corresponda. Por lo tanto, el puntaje W_D aportado por **DPS** se calcula de la siguiente manera:

$$W_D(a_1,b_1) = \frac{\Delta_M - \overline{\Delta}(a_1,b_1)}{\Delta_M}, \tag{3.13}$$

donde $\overline{\Delta}(a_1,b_1)$ es la distancia promedio entre las palabras a_1 y b_1.

A continuación se describe la función de puntaje para la detección de relaciones de oposición y aspectos particulares de los agrupamientos ya descritos.

Cálculo del puntaje aportado por RCL

Las RCL representan el segundo rasgo prioritario en esta identificación, pero su efectividad está en relación de sus características, es decir, si las redes son *reflexivas* o *disyuntivas*. Debido a las hipótesis enunciadas en la sección 2.3 con respecto a los tipos de RCL, es natural pensar que las *RCL reflexivas* aportan una puntuación más estable a la función de puntaje global que las *RCL disyuntivas*, por lo que se decidió tratar los pares relacionados de manera diferente según esas características. Por lo tanto, se formaron dos grupos, dividiendo el conjunto de los pares candidatos en *RCL reflexivas* y *RCL disyuntivas*. Más adelante, en ésta sección, se describe la función de puntaje global con base en estos grupos, para la cuál es necesario primero describir la forma de calcular el puntaje parcial aportado por las RCL.

El cálculo del puntaje parcial $W_{RCL}(a_1,b_1)$, se determina de la siguiente manera:

$$W_{RCL}(a_1, b_1) = \frac{\max\limits_{(a,b)\in G_R}\{s_r(a,b)\} - s_r(a_1, b_1)}{\max\limits_{(a,b)\in G_R}\{\max\limits_{(a,b)\in G_R}\{s_r(a,b)\} - s_r(a_1, b_1)\}} \tag{3.14}$$

donde, $\max\limits_{(a,b)\in G_R}$ representa el valor máximo obtenido por todos los pares relacionados del grupo, con base en la *similitud relativa* s_r. Al complementar, de la *similitud relativa* máxima la *similitud relativa* a cada pareja, se obtiene la *similitud* invertida o complementada, ahora, el máximo de esta *similitud* complementada $\max\limits_{(a,b)\in G_R}\{\max\limits_{(a,b)\in G_R}\{s_r(a,b)\} - s_r(a_1, b_1)\}$ se hace corresponder a uno y los demás puntajes se calculan en proporción directa. Así, $W_{RCL}(a_1, b_1)$ está normalizado entre cero y uno, y el puntaje aportado por las RCL privilegia a los pares cuya s_r es más pequeña.

RCL reflexivas.

El puntaje total $S_g(a_1, b_1)$ es la suma de todos los valores de los rasgos. Este valor se calcula como:

$$S_g(a_1, b_1) = W_{PLS}(a_1, b_1) + W_{DPS}(a_1, b_1) + W_{RCL}(a_1, b_1), \tag{3.15}$$

donde $W_{PLS}(a_1, b_1)$ es el puntaje obtenido por las expresiones regulares que empatan los contextos que contienen tanto a_1 como b_1, $W_{DPS}(a_1, b_1)$ es el puntaje aportado por DPS, y $W_{RCL}(a_1, b_1)$ es el puntaje aportado por las RCL. Cada peso está normalizado en el rango $[0, 1]$. Así, S_g es menor que 3. El cálculo de los valores anteriores fue descrito en C. Lucero et. al. (2004) [6].

Para decidir si un par de términos relacionados, para este agrupamiento, se etiquetan con relación de *oposición*, basta que $S_g \geqslant 1.5$ unidades.

RCL disyuntivas.

El método consiste en tomar en cuenta los rasgos DPS y PLS. Este criterio tuvo su base en la experimentación. Se comprobó que estos dos rasgos juntos, para el

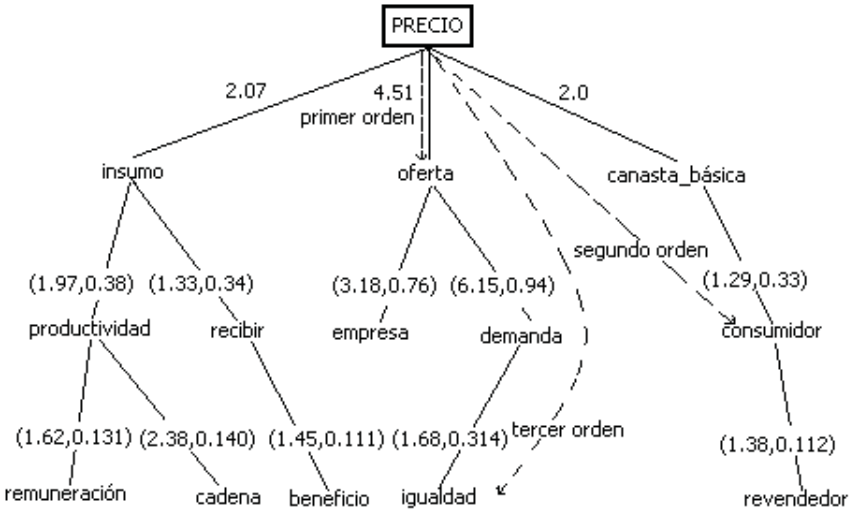

Figura 3.3: Fragmento de la RCL para la palabra *precio*

agrupamiento, dieron mejores resultados que la combinación de los tres rasgos. Así, la función de puntaje cumple, en este caso, $S_g(a_1, b_1) \leq 2$.

Para decidir si un par de términos relacionados, para este agrupamiento, se etiquetan con relación de *oposición*, basta que $S_g \geqslant 1$ unidad.

En la figura 3.4 se presenta de manera esquemática, el proceso de identificación de relaciones de *oposición* descrito anteriormente. En el esquema destacan los principales elementos considerados por el método, como son: la colección de *vecinos cercanos*; el uso de PLS para la selección de los *vecinos cercanos* candidatos a estar en relación de *oposición*; el uso diferenciado de los rasgos en la función de puntaje global, con base en el tipo de RCL de los pares; y el conjunto de pares identificados por el método como opuestos[9].

A continuación se describen los resultados obtenidos en esta etapa y más adelante se presenta un método similar para la identificación de relaciones de *amplitud*.

[9]En este trabajo, relaciones de *oposición* y *opuestos* se utilizan indistintamente.

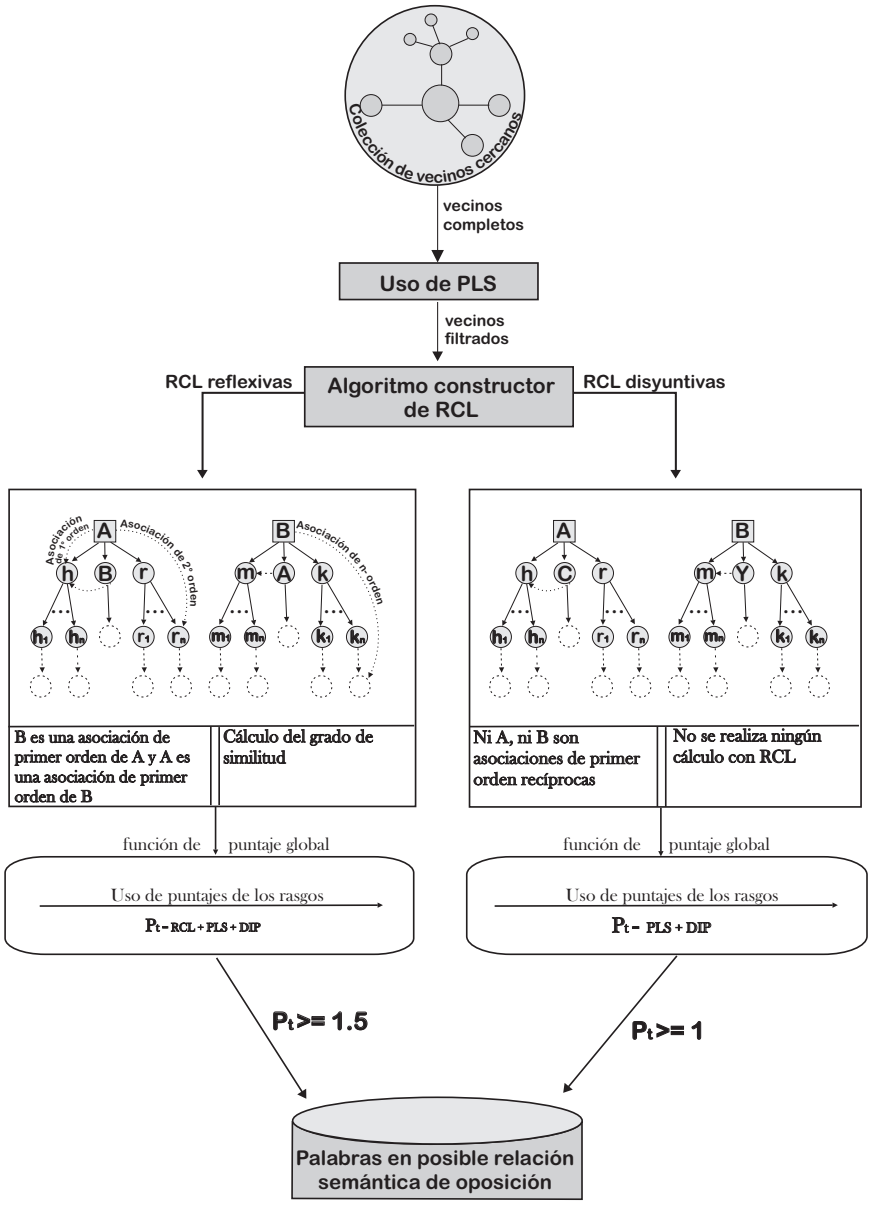

Figura 3.4: Proceso para la identificación de relaciones de *oposición*

Resultados en la identificación de relaciones de *oposición*

De los 65 pares de palabras relacionadas, seleccionadas por las ER, 20 cayeron en el grupo de *RCL reflexivas* y 45 en *RCL disyuntivas*. En las *RCL reflexivas* se marcaron 15 pares en relación de *oposición*, 9 de manera correcta y 6 de manera incorrecta. En las *RCL disyuntivas* se marcaron 33 de los cuales 18 fueron correctos. De manera general, de 65 pares, el método marcó 48 de los cuales 27 fueron marcados de manera correcta, obteniéndose así una precisión de 56.25 %. En la tabla 3.5 se puede observar una muestra de los pares identificados como *opuestos*, y en la tabla A.1 (del apéndice) se presentan los resultados completos.

Palabra A	Palabra B	Relación
Absoluto	**relativo**	Antónimos
Aumento	**disminución**	Antónimos
Barato	**caro**	Antónimos
Consumidor	**productor**	Antónimos
Cualitativo	**cuantitativo**	Antónimos
Demanda	**oferta**	Antónimos
Distrito	**provincia**	Hipónimos
Entrada	**salida**	Antónimos
Escasez	**abundancia**	Antónimos
Explícito	**implícito**	Antónimos
Falso	**verdadero**	Antónimos

Tabla 3.5: Una muestra de pares detectados en relación de *oposición*.

Es importante comentar que muchos de los pares de palabras, marcados de manera incorrecta, se encontraban en relación de *amplitud*, lo que hizo suponer que si lográbamos identificarlos de alguna manera, la precisión aumentaría. Este tópico será visto en la sección 3.3.

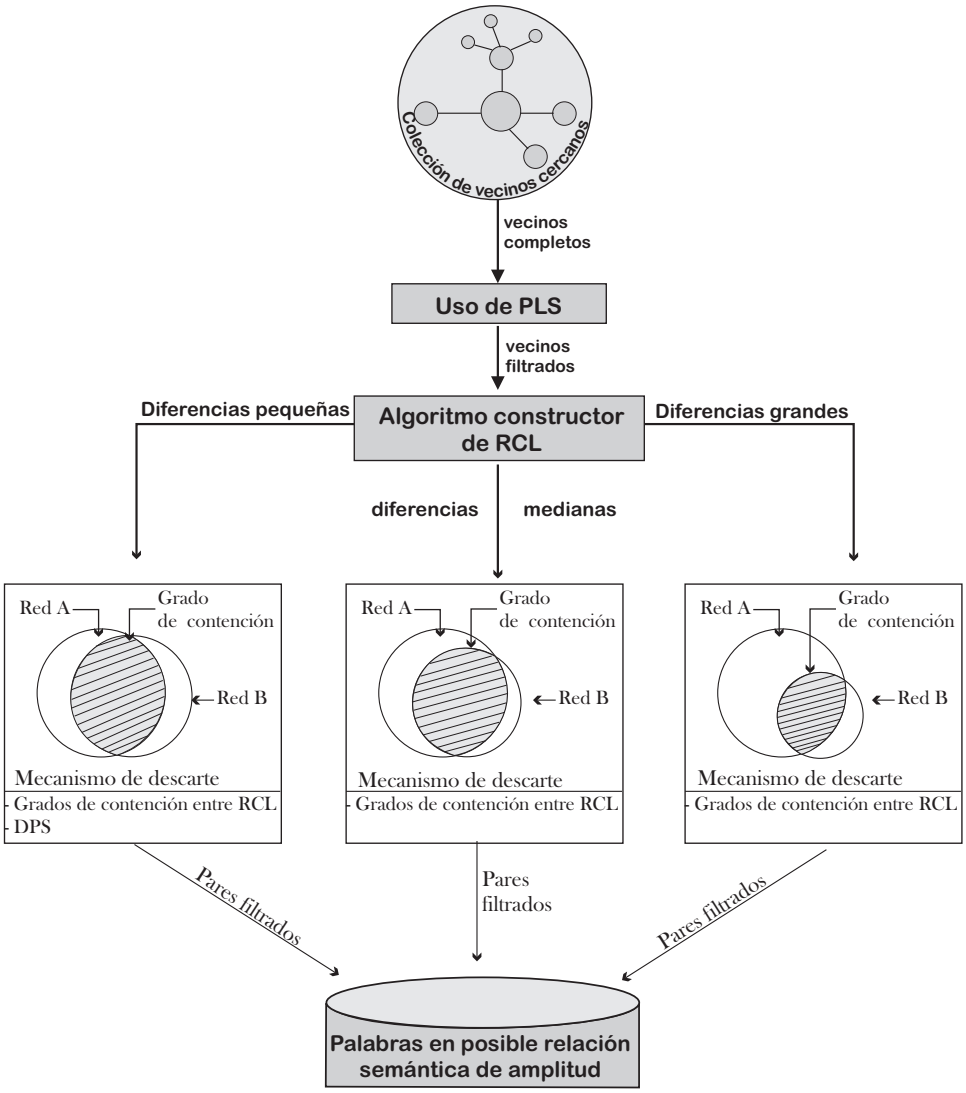

Figura 3.5: Proceso para la identificación de relaciones de *amplitud*

3.2.3. Identificación de relaciones de *amplitud*

En la figura 3.5 se presenta un método para la identificación de relaciones de *amplitud*, para lo cuál, se parte de una colección de palabras relacionadas, estas se filtran mediante el uso de PLS; después se utiliza el grado de *contención* entre las RCL como mecanismo de eliminación, con base en las diferencias de tamaños de las redes.

Para esta tarea, al igual que en la identificación de *opuestos*, los PLS son el rasgo más confiable, por lo que fueron utilizados como mecanismo de selección de pares de palabras relacionadas a ser analizadas. Por medio de los PLS[10] se obtuvo un grupo de 105 pares de los 635 del *thesaurus*, con un umbral de selección $U = 1.2$.

El cálculo del umbral de selección de pares candidatos a estar en relación de *amplitud*, para otros dominios, es similar al usado en relaciones de *oposición* (ver ecuaciones 3.6 y 3.7), con $\sum P_{ER} \approx 470$ y $n = 635$, $f = 1.62$.

Es importante hacer notar que f será usado en la etapa de prueba para establecer U'.

Por otro lado, bajo la hipótesis de que el grado de *contención* y las diferencias de tamaños entre redes pueden señalar algún tipo de relación de *amplitud* (ver sección 2.5), el grupo seleccionado por los PLS fue dividido en tres sub-grupos llamados: **diferencias grandes**, **diferencias medianas**, y **diferencias pequeñas**. El grado de *contención* de las RCL funcionó como mecanismo de eliminación según la agrupación.

Más formalmente: sean Γa_1 y Γb_1 los tamaños, en número de nodos, de las redes a_1 y b_1, con $\delta(a_1, b_1) = |(\Gamma a_1 - \Gamma b_1)|$ la diferencia de tamaños de dichas redes, y sean los grupos $D_G \in T_S$ (diferencias grandes), $D_M \in T_S$ (diferencias medianas) y $D_P \in T_S$ (diferencias pequeñas) con rangos en porcentaje de diferencia de tamaños de redes [67,100], [33.3,67) y [0,33.3), respectivamente. Para determinar a qué grupo

[10] Se usaron 43 PLS para relaciones de *amplitud*.

pertenece cada par, se calcula el porcentaje de la diferencia $\delta'(a_1, b_1)$ de la siguiente manera:

$$\delta'(a_1, b_1) = \frac{\delta(a_1, b_1)}{max\{\delta(a_1, b_1)\},} \tag{3.16}$$

y se verificó a qué rango pertenece.

Los umbrales de eliminación, establecidos para cada grupo son: 20 %, 13 % y 20 % para D_G, D_M y D_P respectivamente. Así, por ejemplo, si un par obtiene un porcentaje de *contención* de 18 % y pertenece D_G, es eliminado, pero si pertenece a D_M, se considera en relación semántica de *amplitud*.

Es importante señalar que la DPS sólo se utilizan en el grupo D_P, debido a que las RCL, en este caso, son más estables, presumiblemente, también las distancias estarán mejor establecidas.

Por último, cabe señalar que los umbrales definidos para eliminar pares, tienen su base en la experimentación. Estos umbrales pueden ser usados como factores para la identificación de nuevos umbrales, como se aprecia a continuación.

Sean $f_0(D)$ cualquiera de los umbrales de eliminación anteriormente citados, y N el número de pares relacionados de cualquier grupo $D \subset \{D_G, D_M D_P\}$, entonces es posible establecer un factor constante $f_1 D$, de la siguiente manera:

$$f_1(D) = \frac{N_D \cdot f_0(D)}{\sum_{i \in D}(\frac{C'_D(a_i, b_i)}{c_i})} \tag{3.17}$$

donde $C'_D(a_i, b_i)$ es el porcentaje de *contención* que existe entre las RCL a_i y b_i en el grupo D, y c_i el número de oraciones del corpus que contienen a ambas: a_i y b_i.

Entonces, para identificar el umbral de eliminación para un dominio diferente U_{Dom}, y según el grupo requerido, se calcula de la siguiente manera:

$$f_0(D) = f_1(D)\frac{\sum_{i \in D}(\frac{C'_D(a_i,b_i)}{c_i})}{N_D} \tag{3.18}$$

considerando N_D, $C'_D(a_i, b_i)$ y c_i dentro del nuevo dominio.

Resultados en la identificación de relaciones de *amplitud*

Los resultados, dependiendo del tipo de agrupamiento, se describen a continuación:

Diferencias grandes. 21 pares cayeron en este grupo. Mediante el grado de *contención* se eliminaron 5 de manera correcta.

Diferencias medianas. 38 pares pertenecen a él. Mediante el grado de *contención* se eliminaron 4 de manera correcta.

Diferencias pequeñas. 45 pares cumplieron con él. Se eliminaron con el grado de *contención* 12 pares, 11 de manera correcta. En éste grupo se utilizó DPS, mediante este rasgo fueron eliminados 22 pares, 18 de manera correcta. El umbral de eliminación de DPS es el límite inferior del rango de los pares en relación de *amplitud*.

Por lo tanto, de los 105 pares de palabras relacionadas, seleccionadas por los PLS, se eliminaron 43, de estos pares 38 fueron eliminados correctamente, quedando 62 pares con 40 en relación de *amplitud* y, por tanto, la precisión alcanzada es 64.5 %. En la tabla 3.6 se presenta una muestra de los resultados obtenidos en esta etapa. Los resultados completos se encuentran en la tabla A.2 (del apéndice).

Es importante comentar que muchos de los pares, marcados de manera incorrecta, se encontraban en relación de oposición, una muestra de estos pares se puede observar en la tabla 3.7

Palabra A	Palabra B	Relación
Amenaza	**Potencial**	TMP
Comisión	**Congreso**	Hipónimos
Debate	**Congreso**	Hipónimos
Matemáticas	**Cálculo**	Hipónimos
Metales_preciosos	**Plata**	Hipónimos
Metal	**Plata**	Hipónimos
Organización	**Persona**	Hipónimos
Oro	**Metal**	Hipónimos
Oro	**Metales_preciosos**	Hipónimos
Productor	**Benefactor**	Hipónimos

Tabla 3.6: Una muestra de pares detectados en relación de *amplitud* .

3.3. Resultados finales de la metodología

En la tabla 3.7 se muestra una porción de los pares de palabras relacionadas que fueron marcados como de *oposición* y de *amplitud* (los resultados completos se pueden observar en la tabla A.3 del apéndice). Si bien, los opuestos pueden ser incluidos en un nivel jerárquico superior (co-hiponimia), no todas las co-hiponimias se pueden ver como relaciones de *oposición*; por ejemplo, las palabras *exportación* e *importación*, marcadas como antónimos, son partes de *actividad-comercial*, es decir: *exportación* e *importación* son hipónimos de *actividad-comercial* y, por tanto, co-hipónimos; los que clasificamos en relación de *amplitud*.

Con base en las observaciones anteriores, hemos decidido despejar, de los resultados de relaciones de *oposición*, los que también fueron identificados como de *amplitud*, e incluir estos pares de la intersección en la relación de *amplitud*. Los resultados finales fueron los siguientes:

Si reubicamos los pares de la intersección (*oposición* y *amplitud*) en solamente el conjunto de relaciones de *amplitud*, entonces, de los 48 pares marcados en relación de *oposición*, 22 estuvieron en la intersección y por tanto fueron removidos, quedando 26 de los cuales 22 son correctos. Así, la precisión obtenida es 84.6 %.

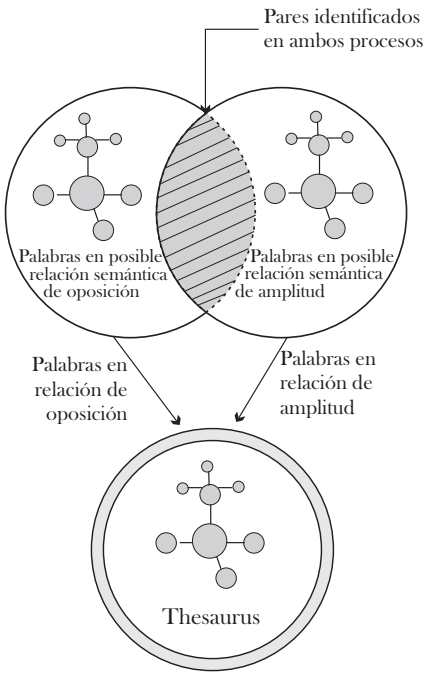

Figura 3.6: Identificación de resultados finales

Al considerar las relaciones de *oposición* como de *amplitud*, en los resultados finales del método para la identificación de relaciones de *amplitud* (sin eliminar los de la intersección), la precisión aumenta de 64.5 % a 79.6 %.

Una parte de los resultados finales se puede ver en la porción de nuestro *thesaurus* de Economía presentado en la tabla 3.8.

Mediante el método descrito, se ha construido un *thesaurus* en el dominio de Economía, constituido por pares de palabras relacionadas con dos tipos de relaciones semánticas, marcadas de manera semi-automática, a saber: de *oposición* y de *amplitud*. En la Tabla 3.8 se muestra una pequeña parte de nuestro recurso, en él se pueden apreciar algunos pares de palabras relacionadas, el tipo de relación semántica (antonimia, sinonimia, co-hiponimia e hiponimia)[11] y nuestra evaluación

[11] El orden de la pareja no fue importante para etiquetar el tipo de relación.

Palabra A	Palabra B	Relación
Alimentación	**Vivienda**	**co-hiponimia**
Azúcar	**Café**	**co-hiponimia**
Castigo	**Premio**	**antonimia**
Exportación	**Importación**	**antonimia**
Legislativo	**Ejecutivo**	**co-hiponimia**
Planta	**Animal**	**co-hiponimia**
Provincia	**Región**	**hiponimia**

Tabla 3.7: Algunos pares de la intersección.

(Evaluación) del tipo de relación (Relación) detectada de manera automática por los algoritmos. En la columna (Evaluación), los calificativos DC y DI indican que el tipo de relación fue detectado de manera correcta o incorrecta respectivamente, y ND indica que no fue detectado ningún tipo de relación semántica para ese par de palabras.

Es importante notar que es posible inferir nuevas relaciones entre los términos del *thesaurus*, incluso pudieran inferirse también algunos tipos de relaciones semánticas, por ejemplo: *pescado* y *aceite* son co-hipónimos porque *pescado* es co-hipónimo de *harina* y esta a su vez es co-hipónimo de *aceite*.

Palabra A	Palabra B	Relación	Evaluación
Países-ricos	**Países-pobres**	co-hiponimia	DC
Clave-Privada	**Clave-Pública**	co-hiponimia	DC
Sur	**Norte**	antonimia	DC
Trabajo	**Producción**	desconocida	ND
Menor	**Mayor**	antonimia	DC
USA	**Japón**	co-hiponimia	DC
Metal	**Oro**	hiponimia	DC
Metal	**Plata**	hiponimia	DC
Oro	**Plata**	co-hiponimia	DI
Oro	**Metal-precioso**	hiponimia	DC
Tradición	**Instinto**	desconocida	DI
Plata	**Metal-precioso**	hiponimia	DC
Oro	**Mina**	desconocida	ND
Plata	**Cobre**	co-hiponimia	DC
Azúcar	**Café**	co-hiponimia	DC
Pescado	**Harina**	co-hiponimia	DC
Aceite	**Harina**	co-hiponimia	DC
Harina	**Trigo**	co-hiponimia	ND
País	**Mercado**	desconocida	ND
Positivo	**Negativo**	antonimia	DC
Absoluto	**Relativo**	antonimia	DC
Organización	**Persona**	hiponimia	DC
Energía	**Tiempo**	desconocida	DI
Falso	**Verdadero**	antonimia	DC
Justicia	**Equidad**	sinonimia	DI
Pobreza	**Población**	desconocida	ND
Teoría	**Hecho**	desconocida	ND
Público	**Privado**	antonimia	DC
Mercancía	**Dinero**	hiponimia	ND
Ingreso	**Gasto**	antonimia	DI
Legistativo	**Ejecutivo**	co-hiponimia	DC
India	**China**	co-hiponimia	DI
Monetario	**Fiscal**	co-hiponimia	DC
Diez	**Veinte**	co-hiponimia	DC
Banco	**Fondos**	desconocida	ND

Tabla 3.8: Fragmento del *thesaurus enriquecido*.

Capítulo 4

Etapa de prueba

Para la construcción del *thesaurus* de Medicina, se ha utilizado un *corpus* sin formato de 5.2 MB de espacio en disco, con un total de 36,000 oraciones y 14329 términos. El procedimiento para la identificación de los pares de palabras relacionadas y los tipos de relaciones semánticas, fue hecho de manera similar al realizado para el *thesaurus* de Economía, con el objeto de verificar su eficiencia en un dominio diferente. Así pues, se construyó un *thesaurus* en el dominio de Medicina, partiendo de 846 pares de palabras relacionadas, a las cuales se les aplicó la metodología de identificación de relaciones de *amplitud* y *oposición*, considerando los umbrales y factores establecidos en la etapa de entrenamiento. A continuación se describe el procedimiento seguido. Más adelante, en la sección 4.2, se presentan los resultados de la aplicación de la metodología al nuevo dominio, y por último, en la sección 4.3 se hace la evaluación de dichos resultados.

4.1. Determinación de umbrales

Como hemos visto en el capítulo anterior, la determinación de un tipo de relación entre dos palabras se realiza calculando un puntaje y, con base en rebasar un umbral, se determina si pertenece a una clase o no. Así, es necesario definir los umbrales para este nuevo dominio, representado por un *corpus*.

La determinación de umbrales relevantes para el dominio, toma en consideración todos los criterios vistos en la etapa de entrenamiento, enseguida, se describen los

elementos particulares del dominio, necesarios para el establecimiento de dichos umbrales, así como los umbrales determinados.

4.1.1. Umbrales para la identificación de relaciones de *oposición*

Para seleccionar el grupo de pares de palabras candidatas a estar en relación de *oposición* se establece U' para el nuevo dominio (ver ecuaciones 3.6 y 3.7), con $n = 846$, $f = 2.92$, y $\sum P_{ER} = 379.5$, de lo que se obtiene $U' = 1.31$. Mediante este umbral se seleccionaron 86 pares candidatos.

Mediante las fórmulas de la sección 3.2.2 se estableció el rango $[0.5, 3]$ de DPS para relaciones de *oposición*. Ahora, para calcular el puntaje aportado por la DPS, la distancia promedio de los 86 pares de palabras seleccionados por los PLS para *opuestos*, las distancias promedio son complementadas con respecto del límite superior del rango de los opuestos ($LS = 3$), y con respecto del máximo; luego se calcula la proporción, tal y como se realizó en la etapa de entrenamiento (ver capítulo 3).

El grupo de parejas seleccionadas por los PLS para el *thesaurus* de Medicina se divide en dos grupos: *RCL reflexivas* y *RCL disyuntivas* con 35 y 51 pares respectivamente. Los puntajes aportados por los PLS se calculan tomando en cuenta la media y la desviación estándar, según el grupo.

Cálculos con *RCL reflexivas.* El promedio de los pesos de los PLS de todos los pares de este grupo es $\overline{x}_R = 3.52$ y la desviación estándar $S_R = 2.57$.

Cálculos con *RCL disyuntivas.* El promedio de los pesos de los PLS de todos los pares de este grupo es $\overline{x}_D = 3.12$ y la desviación estándar $S_D = 1.92$.

4.1.2. Umbrales para la identificación de relaciones de *amplitud*

Para seleccionar el grupo de pares de palabras candidatas a estar en relación de *amplitud* se establece U', con $n = 846$, $f = 1.62$, y $\sum P_{ER} = 605.8$, de lo que se

obtiene $U' = 1.16$. Mediante este umbral se seleccionaron 145 pares candidatos, de los cuales se obtuvieron tres grupos con base en las diferencias de los tamaños de las redes: **Diferencias pequeñas, Diferencias medianas** y, **Diferencias grandes**. Para saber a qué grupo pertenece un par, se identifica el máximo de la diferencia de todos los pares y se hace la proporción (ver sección 3.2.3). Los umbrales de descarte para los diferentes grupos queda de la siguiente manera: $D_P = 33.9\,\%$, $D_M = 27.9\,\%$ y $D_G = 57.9\,\%$; los factores de descarte fueron: $f_1D_P = 8.46$, $f_1D_M = 6.23$ y $f_1D_G = 13.6$.

Por otro lado, mediante las fórmulas de la sección 3.2.2 se estableció el rango $[0.5, 5.13]$ de DPS para relaciones de *amplitud*, y sutiliza $LI = 0.5$ como un mecanismo de descarte en el grupo **Diferencias pequeñas** (ver etapa de entrenamiento).

4.2. Resultados

Una muestra de los resultados finales de la aplicación de la metodología, para la identificación de relaciones de *oposición*, según el grupo (*RCL reflexivas* y *RCL disyuntivas*), se pueden apreciar en las tablas 4.1 y 4.2, respectivamente (los resultados completos se encuentran las tablas B.1 y B.2, del apéndice). En las tablas se pueden observar también, los rasgos utilizados y su aporte a la función de puntaje global, según el tipo de RCL.

Ahora, Una muestra de los resultados de la metodología, para la identificación de relaciones de *amplitud*, para los diferentes grupos, se muestran en las tablas 4.3, 4.4 y 4.5. En dichas tablas, es posible apreciar: los pares relacionados (Palabra A y Palabra B), el número de nodos de las RCL para cada palabra de los pares (N.A y N.B), el número de nodos comunes en ambas redes (N. Comunes), y los porcentajes de *contención* (%Cont), la diferencia en los tamaños de las RCL (%Dif), y la DPS. Los resultados completos se presentan en las tablas B.3, B.4 y B.5, del apéndice.

Palabra A	Palabra B	DIP	PLS	RCL	Global
Total	**Parcial**	0.9804	1.0000	0.7205	2.7009
Gramo	**Negativo**	0.9630	0.3916	0.7131	2.0677
Oral	**Anal**	0.8889	0.2383	0.7956	1.9227
Leve	**Moderado**	0.9167	0.4918	0.6251	2.0335
Estrés	**Ansiedad**	0.8095	0.2635	0.5370	1.6100

Tabla 4.1: Pares detectados como de *oposición* por la metodología, para *RCL reflexivas*.

Palabra A	Palabra B	DIP	PLS	Global
Cocido	**Crudo**	0.9167	1.0000	1.9167
Secreción-vaginal	**Semen**	0.7667	1.0000	1.7667
Fruto	**Vegetal**	0.9333	0.7156	1.6489
Semana	**Mes**	0.3750	0.9941	1.3691
Menor	**Mayor**	0.1008	1.0000	1.1008

Tabla 4.2: Pares detectados como de *oposición* por la metodología, para *RCL disyuntivas*.

Palabra A	Palabra B	N. A	N. B	N. Comunes	%Cont	%Dif
Receptor	**Antagónico**	3172	1290	942	73.02	68.511
Bradicardia	**Hipotensión**	425	2521	404	61.41	76.301

Tabla 4.3: Pares detectados como de *amplitud* por la metodología, para el grupo **Diferencias grandes**.

Para estimar qué tan correctos son los resultados obtenidos en ésta etapa de prueba de la metodología, se buscó la opinión de expertos en el área, para lo cuál se extrajo una muestra y se formuló un cuestionario de opción múltiple, para que el experto eligiera la respuesta que le pareciera más acertada. En la siguiente sección se describe como fue realizada la prueba y los resultados obtenidos.

Palabra A	Palabra B	N. A	N. B	N. Comunes	%Cont	%Dif
Memoria	**Lenguaje**	3024	2041	1321	64.72	35.78
Motor	**Sensitivo**	3109	1318	1147	64.57	65.20
Volumen	**Minuto**	2925	1503	1091	63.87	51.77
Cerdo	**Cisticercos**	1655	2708	1293	61.45	38.33
Célula	**Crecimiento**	3185	1865	1244	60.05	48.05
Queso	**Leche**	1296	2380	1003	59.10	39.46
Suplemento	**Calcio**	1470	2657	851	57.89	43.21
Habla	**Lenguaje**	1171	2217	661	56.45	38.08
Hormona	**Insulina**	2539	1407	763	54.23	41.21
Alergia	**Asma**	2452	1416	1014	53.53	37.71

Tabla 4.4: Pares detectados como de *amplitud* por la metodología, para el grupo **Diferencias medianas**.

Palabra A	Palabra B	DPS	N. A	N. B	N. Comunes	%Cont	%Dif
Abdomen	**Torax**	1.17	2297	1826	1109	60.73	17.15
Ampolla	**Llaga**	1.20	1566	2020	1119	71.46	16.53
Ansiedad	**Depresión**	1.20	2553	1841	1135	61.65	25.92
Aprendizaje	**Escritura**	2.35	2152	1649	1001	60.70	18.31
Auditivo	**Visual**	0.85	1473	2217	992	51.80	27.08
Blanco	**Begro**	1.33	1919	1468	864	50.95	16.42
Carne	**Animal**	1.56	1881	2167	933	49.60	10.41

Tabla 4.5: Algunos pares detectados como de *amplitud* para el grupo **Diferencias pequeñas**.

4.3. Evaluación de resultados

Para evaluar los resultados obtenidos en la etapa de prueba, se tomó una muestra y se le envió a cuatro jueces, quienes habrían de seleccionar el tipo de relación semántica que a su juicio correspondiera[1].

En la muestra se presentaron de manera explicita los pares de palabras relacionadas y las opciones con base en los posibles tipos de relaciones semánticas entre ellas.

[1] Los jueces fueron cuatro médicos.

La clasificación de los tipos de relaciones y los criterios considerados para su identificación, son los siguientes:

parte-de (PD). Es el caso del par (aguja, jeringa), es decir se considera que aguja es parte de jeringa. Así, se espera que la evaluación sea **PD**.

equivalente-a (EQ). Se trata de palabras que al menos en un contexto son equivalentes (se puede usar una o la otra indistintamente), por ejemplo, (medicina, medicamento).

es-un (EU). En este caso la primera palabra pertenece a la clase que denota la segunda. Un ejemplo de ello es: **fémur es un hueso**. Es importante hacer notar que el orden en los pares de las palabras no es importante, es decir: en la lista de palabras relacionadas, por ejemplo, puede aparecer el par (droga, tabaco) o (tabaco, droga), en ambos casos se establece la relación **tabaco es-una droga**.

misma-clase-que (MC). Se consideran con esta relación pares como: (antebrazo, mano); es decir ambos términos pertenecen al mismo grupo (partes del cuerpo).

opuesto-a (OA). Esta relación la cumplen parejas como (sano, enfermo).

ninguna (N). Es posible que una pareja no cumpla ninguna de las anteriores relaciones; esto es, que los términos sí estén relacionados, pero que el tipo de relación no se haya contemplado en ninguna de las que se han definido antes.

sin-relación-con (SR). Esta marca se emplea en los casos que no existe relación entre la pareja. Por ejemplo: (aeropuerto, hijo).

Es importante notar que es posible tener parejas con más de una relación. Por ejemplo, (sano, enfermo), como se ha mencionado, tiene marca OA, pero también en determinado contexto puede considerase la marca MC, debido a que son dos estados de salud. Para estos casos se espera que se elija la marca que el juez considera más

común.

La muestra estuvo constituida de 81 pares de palabras clasificadas por el sistema como de *oposición* y de *amplitud*, y por pares para los que el sistema no identificó ningún tipo de relación semántica. La muestra, evaluada por los jueces, fue analizada, contemplando las respuestas dentro de las siguientes clases de términos con algún tipo de relación: *oposición* (OP), *equivalencia* (EQ), *amplitud* (AM) y *ninguna (N)*. Donde la clase **AM** representa a las opciones (**PD, MC** y **EU**).

A continuación se presentan los criterios considerados para la evaluación de los resultados obtenidos en la etapa de prueba, con base en la evaluación de los jueces, y más adelante, en la sección 4.3.2, se realiza un cálculo del grado de acuerdo de los jueces con respecto a dicha evaluación.

4.3.1. Cálculo del grado de precisión de los resultados

Para cada par de la muestra, si la relación semántica marcada por el sistema es la misma que la marcada por un juez, considerando la clasificación anterior, su puntuación será de una unidad, en cualquier otro caso se tendrá una puntuación de cero unidades: a este criterio se le llama **criterio duro** (CD), esta forma de evaluación también fue usada en H. Salazar (2004) [30], para evaluar extractos de textos obtenidos de manera automática de un corpus sin formato. Ahora, si consideramos que las respuestas de los jueces se encuentran entre clases cercanas; por ejemplo, si el sistema detectó una relación de *amplitud* entre un par y el juez marcó una relación de equivalencia, entonces el puntaje asignado será de $\frac{3}{4}$. Es decir, se considera un puntaje diferente según el grado de cercanía que exista entre los resultados del sistema y la evaluación de los jueces. esta forma de evaluar los resultados, en estadística, se le conoce como **Criterio suave** (CS). En la tabla 4.6 se muestran los puntajes asignados a los pares con base en los resultados del sistema y la evaluación de los jueces.

Sistema	Juez	Puntaje
Amplitud	**Amplitud**	1
Amplitud	**Equivalencia**	$\frac{3}{4}$
Amplitud	**Oposición**	$\frac{3}{4}$
Amplitud	**Ninguna**	$\frac{2}{4}$
Oposición	**Oposición**	1
Oposición	**Amplitud**	$\frac{3}{4}$
Oposición	**Ninguna**	$\frac{2}{4}$
Oposición	**Equivalencia**	$\frac{1}{4}$
Amplitud	**SR**	0
Oposición		

Tabla 4.6: Criterios para los pares de la muestra detectados por el sistema en relación de *oposición* y de *amplitud*.

En la tabla se pueden observar tres aspectos sobresalientes: primero, el puntaje más alto fue asignado a las parejas en las cuales el sistema identificó el mismo tipo de relación semántica marcada por los jueces, tal es el caso de (*amplitud*, *amplitud*); segundo, el puntaje más bajo (1/4) fue asignado a resultados extremos, tal es el caso, de que el sistema haya detectado una relación de *oposición*, y el juez una relación de equivalencia, o viceversa; y tercero, no se asigna ningún puntaje a los pares que el juez marcó como SR y que el sistema detectó como de *oposición* o *amplitud* [2]. Un elemento más a considerar es que los puntajes sean obtenidos mediante una división entre cuatro, esto se debe a que hemos clasificado en cuatro grupos los posibles tipos de relaciones semánticas entre las parejas.

Ahora, Los criterios tomados para la evaluación de los resultados en los que el sistema no detectó algún tipo de relación (SND), se pueden observar en la tabla 4.7, donde destaca el puntaje de $\frac{1}{4}$ asignado a las parejas para las que los jueces marcaron como no relacionadas (SR), esto se debe a que, si bien el sistema no identificó el tipo de relación, no dice nada con respecto a que exista o no relación entre ellas, por lo que

[2]Creemos que ese es el caso más drástico en la evaluación, puesto que se supone que todos los pares tienen algún tipo de relación semántica, aunque desconocida, ya que se han obtenido de un *thesaurus*.

no es tan drástico asignar un puntaje en este caso como en el caso en el que el sistema identifique una relación y los jueces opinen que las palabras de la pareja ni siquiera estén relacionadas.

Sistema	**Juez**	**Puntaje**
No detectada	**Ninguna**	1
No detectada	**Oposición**	$\frac{2}{4}$
No detectada	**Amplitud**	$\frac{2}{4}$
No detectada	**Equivalencia**	$\frac{2}{4}$
No detectada	**SR**	$\frac{1}{4}$

Tabla 4.7: Criterios para los pares de la muestra no detectados por el sistema.

Más formalmente, sean los siguientes elementos de la función $C_j(x, y)$ criterio de evaluación: j representa la evaluación de un juez; s refiere a la evaluación del sistema; cl representa una de las cuatro clases definidas anteriormente, y (x, y) es un par de términos de la muestra.

Entonces, la función de puntuación para el criterio duro $CD_j(x, y)$ es:

$$CD_j(x,y) = \begin{cases} 1 & \text{Si } (cl_j(x,y) = OP \wedge cl_s(x,y) = OP) \vee \\ & (cl_j(x,y) = AM \wedge cl_s(x,y) = AM) \vee \\ & (cl_j(x,y) = N \wedge cl_s(x,y) = N). \\ 0 & \text{En otro caso} \end{cases} \tag{4.1}$$

y la función de puntuación para el criterio suave $CS_j(x, y)$ es:

$$CD_j(x,y) = \begin{cases} 1 & \text{Si } CD_j(x,y) \text{ se cumple.} \\ \frac{3}{4} & \text{Si } (cl_j(x,y) = EQ \wedge cl_s(x,y) = AM) \vee \\ & (cl_j(x,y) = OP \wedge cl_s(x,y) = AM) \vee \\ & (cl_j(x,y) = AM \wedge cl_s(x,y) = OP). \\ \frac{2}{4} & \text{Si } (cl_j(x,y) = N \wedge cl_s(x,y) = AM) \vee \\ & (cl_j(x,y) = N \wedge cl_s(x,y) = OP) \vee \\ & (cl_j(x,y) = EQ \wedge cl_s(x,y) = N) \vee \\ & (cl_j(x,y) = OP \wedge cl_s(x,y) = N) \vee \\ & (cl_j(x,y) = AM \wedge cl_s(x,y) = N). \\ \frac{1}{4} & \text{Si } (cl_j(x,y) = EQ \wedge cl_s(x,y) = OP) \vee \\ & ((cl_j(x,y) = SR \wedge cl_s(x,y) = N) \Leftrightarrow (x,y) \in \text{SND}). \\ 0 & \text{En otro caso} \end{cases} \tag{4.2}$$

Ahora, para calcular el puntaje promedio P_t, con base en la evaluación de los jueces a los pares de la muestra, mediante el criterio suave (CS), se calcula con la ecuación:

$$P_t(x,y) = \frac{1}{N_j} \sum_{j=1}^{N_j} CS_j(x,y), \tag{4.3}$$

donde N_j es el número total de jueces que evalúan la muestra (en nuestro caso $N_j = 4$).

Por otro lado, el valor que indica el grado de precisión G_p de los resultados, según la opinión de los jueces y la utilización del criterio suave como mecanismo de valoración, se calcula con la ecuación 4.3.1, donde $N_{x,y}$ representa al número de parejas de la agrupación que se está evaluando.

$$G_p = \frac{1}{N_{x,y}} \sum_{i=1}^{N_{x,y}} P_t(x_i, y_i) \tag{4.4}$$

La evaluación de los resultados, hecha mediante CS, según los agrupamientos: *oposición, amplitud* y *SND* (no detectados por el sistema) se muestran en las tablas 4.8, 4.9 y 4.10.

Palabra A	**Palabra B**	**J1**	**J2**	**J3**	**J4**	$\overline{x}$	Ac_J
oral	**anal**	0.75	1	0.75	1	0.875	0.333
estrés	**ansiedad**	0.75	0.75	0.75	0.75	0.75	1
(secreción-	**semen**	1	0	0.75	0	0.4375	0.333
vaginal)							
fruta	**verdura**	0.75	0.75	0.5	0.75	0.6875	0.666
fruto	**vegetal**	0.75	0.75	0.5	0.75	0.6875	0.666
blando	**paladar**	0.75	0.75	0.75	0.5	0.6875	0.666
						$G_p \cong 0.687$	$G_a \cong 0.61$

Tabla 4.8: Evaluación de los resultados identificados como de *oposición* por el sistema.

Por último, el grado de precisión total G_pT obtenido mediante la aplicación del CS a la evaluación de los jueces, se puede observar en la tabla 4.11.

4.3.2. Cálculo del grado de acuerdo entre los jueces

El grado de acuerdo entre los jueces, con respecto de la muestra, se calculó de dos maneras: mediante un criterio de coincidencias entre los jueces a las respuestas, y mediante la medida *Kappa*, A. Green (1997) [2]. A continuación se describe el criterio de coincidencias usado y más adelante los resultados obtenidos mediante la aplicación del índice *Kappa*.

Criterio suave

Para calcular el grado de acuerdo total G_aT entre los jueces, mediante el criterio suave, se asignó un puntaje proporcional según el grado de acuerdo $Ac_j(x, y)$ con cada uno de los pares (x, y), como se muestra en la siguiente ecuación:

Palabra A	Palabra B	J1	J2	J3	J4	$\overline{x}$	Ac_J
hormona	**insulina**	0.75	1	0.5	1	0.8125	0.333
dolor	**fiebre**	1	1	0.75	1	0.9375	0.666
genes	**cromosomas**	1	1	1	1	1	1
(dolor-de-cabeza)	**fiebre**	1	1	1	1	1	1
picor	**sensación**	1	1	1	1	1	1
cerebro	**médula-espinal**	1	1	1	1	1	1
grasa	**colesterol**	1	1	1	1	1	1
(relación-sexual)	**anal**	0	1	0.5	1	0.625	0.333
columna	**cadera**	1	1	1	1	1	1
radioterapia	**quimioterapia**	0.75	1	1	1	0.9375	0.666
garganta	**nariz**	1	1	1	0.5	0.875	0.666
ansiedad	**depresión**	1	1	1	1	1	1
abdomen	**torax**	1	1	1	1	1	1
labio	**boca**	1	1	1	1	1	1
cuello	**cabeza**	1	1	1	1	1	1
lumbar	**servical**	1	1	1	1	1	1
sífilis	**gonorrea**	1	1	0.75	1	0.9375	0.666
vitamina	**mineral**	0.75	1	1	1	0.9375	0.666
colon	**recto**	1	1	1	1	1	1
ataque	**(derrame-cerebral)**	0	1	1	0	0.5	0.333
axila	**cuello**	1	0	0	1	0.5	0.333
						$G_p \cong 0.81$	$G_a \cong 0.7$

Tabla 4.9: Evaluación de los resultados identificados como de *amplitud* por el sistema.

$$Ac_J(x,y) = \begin{cases} 1 & \text{Si todos los jueces coinciden.} \\ \frac{2}{3} & \text{Si tres jueces coinciden.} \\ \frac{1}{3} & \text{Si dos jueces coinciden.} \\ 0 & \text{Si ningún juez coincide.} \end{cases} \tag{4.5}$$

donde, un acuerdo total con respecto a un par, esto es: que los cuatro jueces coincidan

Palabra A	Palabra B	J1	J2	J3	J4	$\overline{x}$	Ac_J
diabetes	**insulina**	0.5	0.5	0.5	1	0.625	0.333
bilis	**conducta**	0.25	0.25	0.25	0.25	0.25	1
aspirar	**punción**	1	0.5	0.5	0.5	0.625	0.333
asfixia	**hipoxia**	0.5	0.5	0.5	0.5	0.5	1
aprendizaje	**dislexia**	0.5	0.5	1	0.5	0.625	0.333
apnea	**sueño**	0.5	0.5	0.25	1	0.5625	0.333
anticuerpo	**eritrocitos**	0.5	0.25	1	0.25	0.5	0.333
anticuerpo	**inmune**	0.5	0.5	0.5	0.5	0.5	0.666
						$G_p \cong 0.53$	$G_a \cong 0.62$

Tabla 4.10: Evaluación de los resultados no identificados por el sistema.

Grupo	G_p	G_a	Kappa
Oposición	0.687	0.61	0.06
Amplitud	0.811	0.7	0.21
No detectadas	0.530	0.62	0.24
	$G_pT \cong 0.676$	$G_aT \cong 0.643$	$\hat{\bar{k}} \cong 0.17$

Tabla 4.11: Evaluación final de los resultados.

en su respuesta con la misma clase, tendrá la puntuación más alta, en caso contrario, ningún punto será asignado a esa pareja.

En la columna Ac_J de las tablas: 4.8, 4.9 y 4.10 se puede observar el puntaje asignado, según el acuerdo parcial de los jueces, a cada pareja de la muestra. Así, por ejemplo, para la pareja *(lumbar, servical)* de la tabla 4.9, todos los jueces han coincidido en que existe una relación de *amplitud*.

Por último, el grado de acuerdo total se obtiene promediando los acuerdos parciales, según la clase a la que corresponden los pares, mediante la ecuación:

$$G_aT = \frac{1}{N_m} \sum_{i=1}^{N_m} Ac_J(x_i, y_i), \tag{4.6}$$

donde, N_m es el número total de pares de la muestra. El grado de acuerdo total G_aT

se puede observar en la tabla 4.11.

Índice Kappa

La Kappa estadística ($\hat{k}$) fue propuesta por J. Cohen (1960) [31] para calcular el grado de concordancia entre dos jueces con base en la clasificación supervisada de sus respuestas a una muestra. A través del tiempo ha sufrido algunas variantes que buscan, mejorarla y generalizarla para considerar n-jueces, en A. Green (1997) [2] se presenta una forma generalizada de la Kappa estadística, de acuerdo con la teoría de J. L. Fleiss (1971) [32].

Sea (n) el número total de elementos de una muestra valorada por (m) jueces, donde (m_i) es el número de valoraciones en el $i - esimo$ elemento, y (k) es el número de categorías dentro de las cuales se puede hacer la clasificación.

Se define x_{ij} como el número de valoraciones en el elemento $i(i = 1, \cdots, n)$ dentro de la categoría $j(j = 1, \cdots, k)$, donde

$$m = \sum_{j=1}^{k} x_{ij} \tag{4.7}$$

Para toda i.
Sea $\bar{m}$ el número de valoraciones por elemento, si el número de valoraciones es igual a cada elemento $\bar{m}$, será igual a m.

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} \tag{4.8}$$

Ahora, sea $\bar{p_j}$ la proporción global de valoraciones (acuerdos observados) en la categoría j, y $\hat{k_j}$ el valor de *Kappa* por categoría j, $j = 1, \cdots, k$.

$$\bar{p_j} = \frac{\sum_{i=1}^{n} x_i}{n\bar{m}} \tag{4.9}$$

el valor de $\hat{k_j}$ es:

$$\hat{k_j} = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\bar{p_j}\bar{q_j}}, \tag{4.10}$$

donde $\bar{q_j} = 1 - \bar{p_j}$.

Por último, el valor generalizado de *Kappa* $\hat{\bar{k}}$ puede ser definido como un promedio de pesos de los valores individuales de *Kappa*:

$$\hat{\bar{k}} = \frac{\sum_{j=1}^{k} \bar{p_j}\bar{q_j}\hat{k_j}}{\sum_{j=1}^{k} \bar{p_j}\bar{q_j}} \tag{4.11}$$

Lo cuál es equivalente a:

$$\hat{\bar{k}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p_j}\bar{q_j}}. \tag{4.12}$$

El valor de *Kappa* $\hat{\bar{k}}$ obtenido, con base en la muestra citada anteriormente, se puede observar en la tabla 4.11.

Es importante recordar que la muestra estuvo constituida de tres sub-muestras, es decir, de pares clasificados por el sistema en relación de *oposición*, *amplitud*, y *no detectados*[3]. Ahora, las respuestas de los jueces podían caer en cinco grupos (valoraciones posibles) que son: *oposición*, *amplitud*, *equivalencia*, *ninguna* y *sin relación*. Así, se calculó *Kappa* por separado, es decir, el cálculo del acuerdo entre los jueces se realizó con respecto de cada una de las sub-muestras. Los resultados se pueden observar en la tabla 4.11. Por último, para conocer el acuerdo entre los jueces para la muestra completa, se calculó el promedio de las *Kappas* de las sub-muestras. Esta forma de calcular el índice *Kappa* por separado, hace posible la comparación entre G_p y G_a obtenidos mediante el CS, y *Kappa*.

Crudamente, $\hat{\bar{k}} \cong 0.17$ indica un acuerdo pobre entre los jueces. Sin embargo, creemos que sería interesante incluir al *sistema de clasificación* como un juez más, y calcular *Kappa* entre los jueces y entre los jueces y el sistema, para lo cuál fue necesario reducir

[3]Pares que el sistema no clasificó.

el número de clases a las que el sistema clasifica (nuevas valoraciones posibles), que son: *oposición*, *amplitud* y *no detectadas*, en esta última clase se han incluido las clases: *equivalencia*, *ninguna* y *sin relación*, y el cálculo de *Kappa* se realizó a la muestra completa[4]. Esto hace congruente la evaluación entre los jueces y el sistema.

Los resultados de esta nueva evaluación fueron los siguientes: $\hat{\bar{k}} = 0.22$ para el grado de acuerdo entre los jueces, y $\hat{\bar{k}} = 0.133$ para el grado de acuerdo entre los jueces y el *sistema de clasificación*. Notar que el grado de acuerdo entre los jueces para este nuevo cálculo de *Kappa* es mayor que el obtenido de manera separada. Dicho aumento era esperado, puesto que el número de valoraciones posibles es diferente en ambos cálculos. Ahora, si suponemos que los jueces tienen la razón (que están de acuerdo), entonces el grado de acuerdo entre ellos y el *sistema de clasificación* es de 60.5 %.

Por otro lado, si seleccionamos las respuestas de la muestra de manera aleatoria (*sistema aleatorio*), y se vuelve a calcular *Kappa* entre los jueces y el *sistema aleatorio*, y entre el *sistema de clasificación* y el *sistema aleatorio*, se obtiene: $\hat{\bar{k}} \cong -0.25$ y $\hat{\bar{k}} \cong -0.14$ respectivamente. Los resultados anteriores, lejos de indicar un grado de acuerdo, indican un grado de desacuerdo[5]. Algunos valores interesantes de *Kappa* son: $\hat{\bar{k}} = 1$ indica un acuerdo perfecto, $\hat{\bar{k}} = 0$ indica un acuerdo o desacuerdo aleatorio[6], y $\hat{\bar{k}} > 0$ ya indica un cierto acuerdo.

Por lo tanto, los resultados de las evaluaciones utilizando *Kappa* estadística nos parecen aceptables, mientras que los resultados de la evaluación y del grado de acuerdo entre los jueces por el criterio suave, en términos generales, nos parecen de aceptables a buenos, sin embargo, es importante seguir investigando en esta línea para mejorarlos.

[4]Se agrupó una sola muestra con las tres sub-muestras ya descritas.

[5]El acuerdo observado es menor que el esperado.

[6]Que no hay acuerdo ni desacuerdo.

Capítulo 5

Conclusiones

Los principales aportes de este trabajo son los siguientes:

1. **Un método para identificar relaciones semánticas.**

   En este trabajo se ha presentado un nuevo método para la identificación de relaciones de *oposición* y de *amplitud*, para lo que se hizo uso de los rasgos: patrones léxico-sintácticos, redes de co-ocurrencia léxica y el patrón distancia promedio de separación.

2. **Estudio de RCL.**

   Se realizó un estudio profundo de las RCL buscando una mejor representatividad de las asociaciones de orden n. Así, Se identificaron algunas características, de las RCL, que fueron importantes en el proceso de identificación de relaciones semánticas, y se calcularon medidas de similitud y contención mediante análisis por niveles de asociación. El estudio sobre RCL tuvo dos vertientes principales: caracterización y refinamiento.

   Caracterización de RCL.

   Para la identificación de relaciones de *oposición* se agruparon los pares candidatos, dependiendo del tipo de RCL que representan: *RCL reflexivas* o *RCL disyuntivas*, y se utilizaron los rasgos como puntajes parciales de una función de puntaje global, según el tipo de agrupación que corresponde. Para la identificación de relaciones de *amplitud* los agrupamientos fueron hechos

considerando las diferencias de tamaños de las RCL, y los rasgos fueron usados como una estrategia de eliminación de parejas con bajos puntajes.

Refinamiento de RCL.

Las RCL fueron refinadas considerando en su construcción una mejor selección de su contenido con base en la medida de IM, y fueron analizadas de forma un tanto distinta a la idea original de P. Edmonds, esto es: las medidas de *similitud* y *contención*, obtenidas de la comparación de las RCL fueron calculadas considerando niveles de asociación comunes, y se usaron criterios diferentes para las *RCL reflexivas* y *RCL disyuntivas*.

3. **Metodología para la creación de thesauri enriquecidos.**

   Se creó una nueva metodología para la construcción de *thesauri* enriquecidos para lo cuál se utilizó la técnica de ventanas de Grefenstette para la detección de relaciones léxicas, y el método aquí propuesto para la identificación de relaciones semánticas. Dicha metodología fue construida y entrenada en el dominio de Economía, y fue probada en el dominio de Medicina. Los resultados obtenidos en la etapa de prueba fueron evaluados considerando la opinión de expertos en el área.

4. **Un primer acercamiento en la identificación de sinónimos.**

   Se ha presentado un nuevo acercamiento en la identificación de relaciones de sinonimia, con miras a establecer una metodología similar a la aquí propuesta, que nos ayude en esta identificación. Para lo que pretendemos utilizar la WEB, y así obtener información relevante con palabras sinónimas.

En suma:

— Los resultados de la evaluación pueden considerarse aceptables, puesto que no dependen de ningún entendimiento previo del dominio, y que los resultados fueron obtenidos automáticamente. Sin embargo, creemos que es posible mejorarlos,

refinando la metodología, identificando nuevos patrones, e investigando en dominios más ricos, como la WEB.

— Los rasgos usados en trabajos previos a éste (patrones léxico-sintácticos), se complementan con RCL y DPS y, por tanto, mejoran la identificación de relaciones semánticas.

— Las RCL encierran información relevante para el establecimiento de relaciones semánticas.

Capítulo 6

Perspectivas

Los trabajos futuros en esta línea de investigación, motivados por el gran interés que existe actualmente en la identificación de relaciones léxico-semánticas para apoyar las tares de PLN, apuntan a la construcción de una metodología para la creación automática de thesauri enriquecidos con la identificación de algunos tipos de relaciones léxico-semánticas. En la figura 6.1 se muestran, a grandes rasgos, los elementos fundamentales de la arquitectura que se plantea.

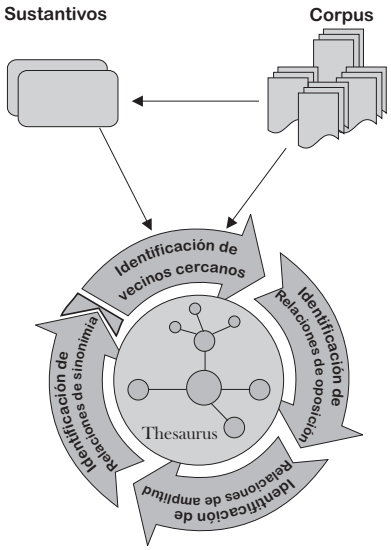

Figura 6.1: Nueva arquitectura para la construcción de thesauri

Una pretensión para lograr esta tarea es utilizar un corpus de dominio específico para la identificación de palabras relacionadas, y la WEB para apoyar a la identificación de los tipos de relaciones. Se piensa aprovechar la gran diversidad de formas de expresión

escrita que en ella existen, para la identificación de nuevos patrones léxico-sintácticos, establecimiento de umbrales independientes del contexto, y para investigar nuevos mecanismos para la detección de relaciones.

En la siguiente sección, se propone un posible método para la identificación de relaciones de sinonimia a partir de un thesaurus, y en la sección 6.2, se plantean algunas pautas para concluir con la construcción de thesauri enriquecidos.

6.1. Identificación de relaciones de sinonimia

La identificación de relaciones de sinonimia en textos crudos, es una tarea difícil de resolver, debido a la existencia de una gran ambigüedad terminológica y a la baja frecuencia con que las palabras sinónimas co-ocurren en los contextos, sin embargo, creemos que existen algunos patrones[1] que pudieran ser descubiertos para identificar algunos sinónimos, y que conjuntamente con el uso de otros rasgos mejorar la cobertura. En la tabla 6.1 se muestran dos expresiones regulares para este propósito, cuyos pesos fueron asignados de la misma forma que en las ER para la identificación de relaciones de *amplitud* y *oposición*. Así, los pesos de 14 y 16 indican que las ER tuvieron una precisión de 70 % y 80 % respectivamente. Estas ER en el dominio de Economía, empatan en contextos como: "ambos: costo y precio son lo *mismo* para el comerciante" (ER 1) y "costo *es lo mismo* que precio, en el mercado" (ER 2). A pesar de que estas ER tienen pesos altos, es necesario comprobar su utilidad en otros dominios y descubrir nuevos patrones; para lo cuál, pensamos que el uso de la WEB, como recurso de información léxico-sintáctica, ligüística y temática, puede ayudar sustancialmente a la solución de este problema.

Una vez que se hayan determinado suficientes ER para la identificación de sinónimos, se puede utilizar una metodología similar a la utilizada en la sección 3.2.2 para la identificación de relaciones semánticas de *oposición*, esto es: a partir de un conjunto de pares relacionados, seleccionar los que tengan mayor peso con base en sus ER,

[1] De muy baja evocación.

Nr	Expresión Regular	Peso
1	Sin word{0,1} **y** word{0,1} Sin **son** word{0,2} [**mismo** | **iguales**]	14
2	Sin word{0,5} [**significa lo mismo** | **se caracteriza como** | **es un sinónimo de** | **es lo mismo que** | **es equivalente a** | ...] word{0,5} Sin	16

Tabla 6.1: ER para sinónimos y sus pesos.

y posteriormente utilizar los tres rasgos descritos anteriormente (DPS, PLS y RCL) en una función de puntaje, pero dentro del grupo de *RCL disyuntivas*: en algunos experimentos realizados con este fin, se observó que la mayoría de las palabras en relación de sinonimia representan este tipo de RCL. En la figura 6.2 se puede apreciar, de manera panorámica, el proceso de identificación de sinónimos que se propone.

6.2. Discriminar entre relaciones semánticas de las clases

Una hipótesis de este trabajo consistía en la idea de utilizar el grado de *similitud* entre RCL para identificar sinónimos, sin embargo, las RCL por si solas no fueron capaces de discriminar entre estas y otras relaciones, por lo que se optó por utilizar otros rasgos que junto con RCL pudieran identificar relaciones de *oposición* y de *amplitud* de manera separada, y lograr así despejar los sinónimos. Este enfoque, aunque mejoró la detección de sinónimos tampoco fue suficiente. De aquí que se haya pensado en la construcción de una metodología para la identificación de sinonimias, para ser aplicada después de la identificación de opuestos y amplios[2]. En un primer experimento con estas ideas, se identificaron 8 pares de palabras, de 18 posibles, que al menos en un contexto se encuentran en relación de sinonimia; la precisión estuvo alrededor del 30 %. Aunque los resultados en este rubro sean aún bastante modestos, creemos que es posible mejorarlos si se revisa la metodología y se identifican nuevos rasgos. En la tabla 6.2 se muestran los 12 pares de palabras relacionadas que lograron el mayor puntaje; los puntajes aportados por los rasgos (RCL, DPS y PLS), y el tipo

[2]En este trabajo, relaciones de *amplitud* y *amplios* se utilizan indistintamente.

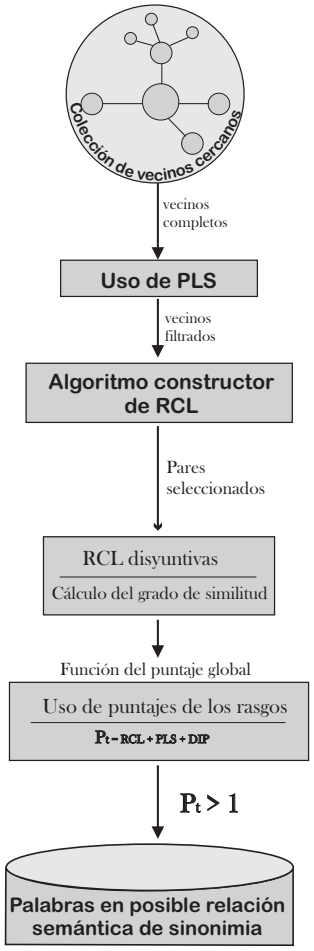

Figura 6.2: Proceso para la identificación de sinónimos

de relación (Relación) que existe entre los pares. Es importante señalar que, para la cuenta global se han despejado los pares que fueron identificados por las metodologías de opuestos y amplios.

Por otro lado, debido a que las relaciones de *oposición* y de *amplitud* han sido consideradas como clases de otros tipos de relaciones más específicas, y debido a la gran dificultad que existe en discriminar estas sub-relaciones, por ahora sólo queda realizar esta tarea de manera semi-automática (ver fig. 6.3) y poder así construir thesauri enriquecidos.

Palabra A	Palabra B	RCL	DIP	PLS	Global	Relación
Inversión	**Gasto**	0,994	0,695	1,000	2,689	sinonimia
Global	**Mundial**	0,529	0,927	1,000	2,456	sinonimia
Recaudación	**Evasión**	0,690	1,000	0,697	2,387	desconocida
Inversión	**Ahorro**	0,802	0,726	0,829	2,357	antonimia
Derecho	**Información**	0,993	0,759	0,599	2,350	desconocida
Digital	**Certificado**	0,903	0,694	0,697	2,294	desconocida
Ciclo	**Etapa**	0,946	0,938	0,363	2,247	sinonimia
Colocación	**Captación**	0,922	0,713	0,607	2,243	desconocida
Valor	**Precio**	0,463	0,938	0,808	2,209	sinonimia
Competente	**Monopolio**	0,761	1,000	0,398	2,159	desconocida
Ciencia	**Conocimiento**	0,592	0,549	1,000	2,141	sinonimia
Dinero	**Valor**	0,734	0,923	0,449	2,105	hiponimia

Tabla 6.2: Muestra de pares de palabras detectadas en relación de sinonimia.

Por último, es importante señalar que a pesar de la dificultad que encierra la tarea de identificar automáticamente relaciones semánticas, es deseable poder discriminar entre las relaciones agrupadas en las clases usadas (*oposición* y *amplitud*). Así, por ejemplo, lograr determinar relaciones de "holonimia", "hiponimia", "hiperonimia", "co-hiponimia" y "co-holonimia"; todas ellas agrupadas dentro de la clase "relaciones de *amplitud*".

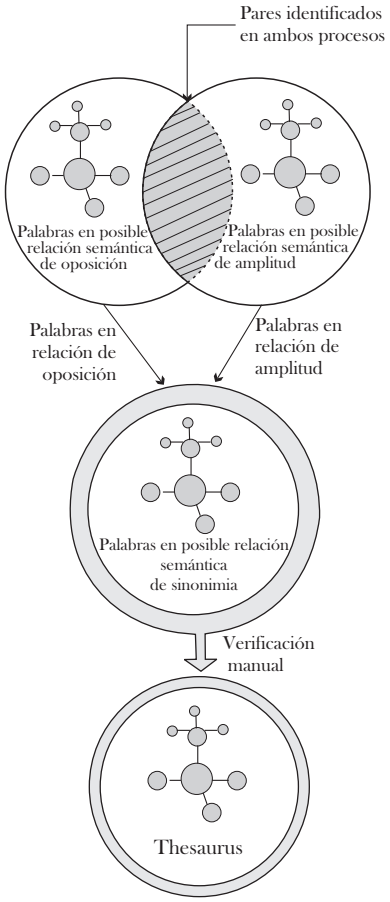

Figura 6.3: Discriminación entre tipos de relaciones léxico-semánticas

Bibliografía

[1] A. Collins & E. Loftus.: "A Spreading-Activation Theory of Semantic Processing", en *Psychological Review, 82(6)*, pp. 407-428, 1975.

[2] A. Green: "Kappa Statistics for Multiple Raters Using Categorical Classifications", en *Proceedings of the 22nd Annual SAS User Group International Conference*, (1997).

[3] A. Pancardo-Rodríguez, M. Montes-y-Gómez, P. Rosso, D. Bucaldi & L. Villaseñor-Pineda.: "Desambiguación Léxica de Sustantivos usando la Web", en *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pp 118-122, Copyright, IBERAMIA 2004. (ISBN 968-863-786-6).

[4] A. Zazo, C. Figuerola, J. Alonso & R. Gómez: "Recuperación de información utilizando el modelo vectorial", en *Taller CLEF-2001*, Departamento de Informática y Automática, Universidad de Salamanca, pp. 1-36, mayo 2002.

[5] C. J. Van Rijsbergen.: "Information Retrieval", *University of Glasgow*, pp. 114-117. Second Edition, 1999.

[6] C. Lucero, D. Pinto & H. Jiménez-Salazar.: "Identificación de antónimos en textos planos", en *Cuarto encuentro de computación*, pp. 203 - 211, Colima-México, CA, Septiembre 2004.

[7] C. Lucero, D. Pinto & H. Jiménez-Salazar.: "A Tool for Automatic Detection of Antonymy Relations", en *Taller de Herramientas y Recursos Lingüisticos para*

el Español y el Portugués, pp 273-281, Copyright, IBERAMIA 2004. (ISBN 968-863-786-6).

[8] C. Lucero, D. Pinto & H. Jiménez-Salazar.: "Una Metodología para la Creación de Thesauri", en *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pp 205-211, Copyright, IBERAMIA 2004. (ISBN 968-863-786-6).

[9] C. Lucero, D. Pinto & H. Jiménez-Salazar.: "Un Método para la Identificación Automática de Relaciones Léxico-Semánticas a partir de un Thesaurus", en *Segundo Congreso Nacional de Ciencias de la Computación*, FCC-BUAP, Noviembre de 2004, Puebla-México.

[10] C. Varaschin & V. Strube de Lima: "Experiments on Extracting Semantic Relations from Syntactic Relations", en *CiCLing 2003*, LNCS 2588, pp. 314-324, 2003.

[11] D. A. Cruse.: "Lexical Semantics", *Cambridge Textbooks in Linguistics*, B. Comrie, C. J. Fillmore, R. Huddleston, R. Lass, D. Lightfoot, J. Lyons, P. H. Matthews, R. Posner, S. Romaine, N. V. Smith & N. Vincent (ed.), Cambridge University Press, 1986.

[12] D. Hindle.: "Noun Classification from Predicate-Argument Structures", in *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990, pp. 268-275.

[13] D. Lin, S. Zhao, L. Qin & M. Zhou.: "Identifying Synonyms among Distributionally Similar Words"en *Proceedings of the Eighteenth International Joint Conferences on Artificial Intelligence (IJCAI-03)*, pp.1492-1493. 2003.

[14] D. McCarthy, R. Koeling, J. Weeds & J. Carroll.: "Finding Predominant Word Senses in Untagged Text", en *42nd Annual Meeting of the Association for Computational Linguistics*, pp. 279-286, Barcelona-España, Julio 2004.

[15] D. Schwab, M. Lafourcade & V. Prince.: "Antonymy and Conceptual Vectors", en *the Proceedings of the 19th Conference on Computational Linguistics*, 2002, pp. 904-910.

[16] D. Walker, R. Amsler.: "The Use of Machine-Readable Dictionaries is Sublanguage Analysis", en *Analysing Language in Restricted domains*, R. Grishman y R. Kittredge (Eds.), Lawrence Erlbaum, Hillsdale, NJ, pp. 69-84. 1986.

[17] E. F. Carcedo.: "Los géneros y su práctica, con una guía gramatical", en *Benemérita Universidad Autónoma de Puebla, Dirección General de Fomento Editorial*, ISBN: 968-863-623-1, pp. 35-38, 2003.

[18] E. Yamamoto, K. Kanzaki & H. Isahara.: "Hierarchy Extraction based on Inclusion Appearance", en *42nd Annual Meeting of the Association for Computational Linguistics*, Interactive Posters/Demostrations Session, pp. 150-153, Barcelona-España, Julio 2004.

[19] G. Grefenstette.: "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches", en *Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, OH, June 1993.

[20] G. Grefenstette.: "Explorations in Automatic *thesaurus* Discovery", Kluwer Academic Publishers, Boston Hardbound, ISBN 0-7923-9468-2 July 1994.

[21] G. Grefenstette.: "Automatic thesaurus generation from raw text using knowledge-poor techniques", *Xerox Report*, 1996.

[22] G. Rigau.: "Resolución Automática de la Ambigüedad Semántica de las Palabras", en *Tecnologías del texto y del habla*, A. Martí & J. Llisteri (Eds.), Universitat de Barcelona, pp. 57-88, 2004.

[23] G. Ruge.: "Experiments on Linguistically Based Term Associations", en *RIAO'91*, pp. 528-545, Barcelona-España. CID, Paris. 1991.

[24] G. Ruge.: "Combining *corpus* linguistics and human memory models for automatic tem association", en *AI Group. Institut fur Informatik*, Munchen. 1999.

[25] G. Salton, A. Wond & C. S. Yang.: "A Vector Space Model for Automatic Indexing", en *Communications of the ACM*, pp. 613-620, Noviembre de 1975.

[26] G. Salton.: "Automatic Term Class Construction Using Relevance", en *A summery of Work in Automatic Pseudoclassification*, Information Processing Management, 16, pp. 1-15, 1980.

[27] G. W. Furnas, T. K. Landauer, L.M. Gómez & S. T. Dumais.: "The vocabukary problem in human-system communication", en *Communications of the ACM 30.964-971*, 1987.

[28] H. Jiménez-Salazar.: "Grado de pertenencia a un dominio y métodos de clasificación", en *Tesis de doctorado*, pp. 1-7, septiembre 2000.

[29] H. Jiménez-Salazar.: "A Method of Automatic Detection of Lexical Relationships Using a Raw *corpus*", en *CiCLing 2003*, LNCS 2588, pp. 325-328, 2003.

[30] H. Salazar: "Obtención de Extractos de Textos con Base en un Corpus", en *Tesis Profesional de Maestría en Ciencias de la Computación*, FCC-BUAP, Mayo de 2004.

[31] J. Cohen.: "A Coefficient of Agreement for Nominal Scales", *Educ. Psychol. Meas*, 20, pp. 37-46, 1960.

[32] J. L. Fleiss.: "Measuring Nominal Scale Agreement Among Many Raters", *Psychol. Bull*, 76, pp. 378-382, 1971.

[33] J. Morato, M. A. Marzal, J. Lloréns & J. Moreiro.: "WordNet Applications", en *Proceedings GWC 2004*, P. Sojka, K. Pala, P. Smrc, C. Fellbaum & P. Vossen (Eds.), pp. 270-278, 2004.

[34] J. Sparck.: "Synonymy and Semantic Classification", en *PhD thesis delivered by University of Cambridge*, Edinburgh: Edinburgh University Press, 1964.

[35] K. Church, W. Gale, D. Hindle & R. Moon: "Lexical Substitutability", en *Computational Approaches to the Lexicon*, Atkins, B. T. S. Atkins and A. ZampoUi (Eds.), Oxford University Press, pages 153-177. 1994.

[36] L. Wanner.: "Lexical Functions in Lexicography and Natural Language Processing", *John Benjamins Publishing Company*, 1996.

[37] M. Hearst.: "Automatic acquisition of hyponyms from large text corpora", en *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992.

[38] M. Hearst.: "Automated Discovery of WordNet Relations", en *WordNet and Electronic Lexical Database*, C. Fellbaum (Ed.), The MIT Press, pp. 131-152. 1999.

[39] M. Lesk.: "Word-Word Associations in Document Retrieval Systems", *American Documentation*, 1, pp. 27-38, 1969.

[40] M. Lesk.: "Automatic sense disambiguation: how to tell a pine cone from an ice cream cone", en *Proceeding of the SIGDOC Conference*, Association for Computing Machinery: New York, 1986.

[41] M. R. Fano.: "Transmission of Information: A Statistical Theory of Communications", *MIT Press, Cambridge, MA.*, 1961.

[42] M. Sanderson & B. Croft.: "Deriving concept hierarchies from text", en *Proceedings of the 22 a Annual lnt. ACM S1G1R Conference on Research and Development in Information Retrieval*, pp. 206 - 213, Berkeley, CA, August 1999.

[43] P. Edmonds.: "Choosing the word most typical in context using a lexical co-ocurrence network", en *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, pp. 507 - 509, 1997.

[44] P. Lewis, P. Baxendale & J. Bennett.: "Statistical Discrimination of the synonymy/Antonymy Relationship Between Words", en *Journal of the ACM*, 14(1), pp. 20-44, 1967.

[45] P. Proctor.: "Longman Dictionary of Contemporary English", *London: Longman*. Ed. (1978).

[46] P. Rosso, F. Masulli, D. Buscaldi, F. Pla & A. Molina.: "Automatic Noun Sense Disambiguation", en *CICLing 2003*: pp. 273-276.

[47] R. Mihalcea.: "Making Sense Out of the Web", en *Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pp 112-117, Copyright, IBERAMIA 2004. (ISBN 968-863-786-6).

[48] S. Banerjee & T. Pedersen.: "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet", en *CICLing 2002*, A. Gelbukh (Ed.), LNCS 2276, pp. 136-145, 2002.

[49] T. Strzalkowski.: "Natural Language Information Retrieval", *Information Processing Mangement*, 31(3), PP. 397-417, 1995.

[50] Y. De Castilla, E. Moyotl, C. Lucero, V. Cortés V, H. Jiménez-Salazar, E. Mendoza, S. Paniagua, D. Pinto & B. Reyes.: "Resultados del Laboratorio de Recuperación de Información de la FCC", en *Primer Congreso Nacional de Ciencias de la Computación*, PP. 87-90, FCC-BUAP, Noviembre de 2003, Puebla-México.

[51] Y. Wilks.: "An intelligent analyzer and understander of english", Communications of the ACM 18. pp. 264-274, 1975.

[52] Y. Wilks & R. Catizone.: "Lexical Tuning", en *CICLing 2002*, pp. 106-125, 2002.

[53] Z. S. Harris.: "Mathematical Structures of Language". Wiley, New York, 1968.

Apéndice A

Resultados de la etapa de entrenamiento

Palabra A	Palabra B	Relación
Absoluto	**Relativo**	Antónimos
Aumento	**Disminución**	Antónimos
Barato	**Caro**	Antónimos
Consumidor	**Productor**	Antónimos
Cualitativo	**Cuantitativo**	Antónimos
Demanda	**Oferta**	Antónimos
Distrito	**Provincia**	Hipónimos
Entrada	**Salida**	Antónimos
Escasez	**Abundancia**	Antónimos
Explícito	**Implícito**	Antónimos
Falso	**Verdadero**	Antónimos
Individual	**Colectivo**	Antónimos
Individual	**Social**	Antónimos
India	**China**	Co-hipónimos
Justo	**Equitativo**	Sinónimos
Menor	**Mayor**	Antónimos
Natural	**Artificial**	Antónimos
Oro	**Plata**	Co-hipónimos
Pequeño	**Grande**	Antónimos
Positivo	**Negativo**	Antónimos
Proveedor	**Cliente**	Antónimos
Público	**Privado**	Antónimos
Sur	**Norte**	Antónimos
Vegetal	**Animal**	Antónimos
Vendedor	**Comprador**	Antónimos
Vender	**Comprar**	Antónimos

Tabla A.1: Pares detectados en relación de oposición.

Palabra A	Palabra B	Relación
Amenaza	**Potencial**	TMP
Comisión	**Congreso**	Hipónimos
Debate	**Congreso**	Hipónimos
Matemáticas	**Cálculo**	Hipónimos
Metales_preciosos	**Plata**	Hipónimos
Metal	**Plata**	Hipónimos
Organización	**Persona**	Hipónimos
Oro	**Metal**	Hipónimos
Oro	**Metales_preciosos**	Hipónimos
Productor	**Benefactor**	Hipónimos
Agricultura	**Industria**	Co-hipónimos
Clave_privada	**Clave_pública**	Co-hipónimos
Cobre	**Plata**	Co-hipónimos
Definición	**Concepto**	Co-hipónimos
Diez	**Veinte**	Co-hipónimos
Educación	**Salud**	Co-hipónimos
Educación	**Vivienda**	Co-hipónimos
Empírico	**Hipótesis**	Co-hipónimos
Estados_unidos	**Japón**	Co-hipónimos
Frances	**Ingles**	Co-hipónimos
Instinto	**Racional**	Co-hipónimos
Millón	**Mil**	Co-hipónimos
Monetario	**Fiscal**	Co-hipónimos
Países_ricos	**Países_pobres**	Co-hipónimos
Presidente	**Congreso**	Co-hipónimos
Producto_liquido	**Producto_bruto**	Co-hipónimos
Social	**Politico**	Co-hipónimos
Vivienda	**Salud**	Co-hipónimos
Empírico	**Teórico**	Antónimos
Ingreso	**Gasto**	Antónimos
Caza	**Animal**	SR
Congreso	**Proyecto**	SR
Industria	**Comercio**	Co-hipónimos
Instinto	**Tradicional**	SR
Instrumento	**Derecho**	SR
Juicio_valor	**Normativa**	SR
Necesidad	**Recursos**	SR
Pensamiento	**Escuela**	SR
Préstamo	**Crédito**	Co-hipónimos
Tradición	**Instinto**	SR

Tabla A.2: Pares detectados en relación de amplitud.

Palabra A	Palabra B	Relación
Acreedor	**Deudor**	Antónimos
Africa	**Sur**	Hipónimos
Alimentación	**Vivienda**	Co-hipónimos
Azúcar	**Café**	Co-hipónimos
Capricho	**Gusto**	Antónimos
Castigo	**Premio**	Antónimos
Cientos	**Miles**	Co-hipónimos
Declarar	**Pacto**	SR
Desempleo	**Inflación**	Co-hipónimos
Empresa	**Persona**	Hipónimos
Energía	**Mina**	SR
Energía	**Tiempo**	SR
Exportación	**Importación**	Antónimos
Futuro	**Presente**	Co-hipónimos
Harina	**Aceite**	Co-hipónimos
Interno	**Externo**	Antónimos
Legislativo	**Ejecutivo**	Co-hipónimos
Pescado	**Harina**	Co-hipónimos
Planta	**Animal**	Co-hipónimos
Provincial	**Región**	Hipónimos
Regional	**Local**	Co-hipónimos
Vino	**Paño**	SR

Tabla A.3: Pares detectados como de oposición y de amplitud.

Apéndice B

Resultados de la etapa de prueba

Palabra A	Palabra B	DIP	PLS	RCL	Global
Total	**Parcial**	0.9804	1.0000	0.7205	2.7009
Gramo	**Negativo**	0.9630	0.3916	0.7131	2.0677
Oral	**Anal**	0.8889	0.2383	0.7956	1.9227
Leve	**Moderado**	0.9167	0.4918	0.6251	2.0335
Estrés	**Ansiedad**	0.8095	0.2635	0.5370	1.6100
Secreción	**Nasal**	0.7436	0.2762	0.5309	1.5506

Tabla B.1: Pares detectados como de *oposición*, para *RCL reflexivas*.

Palabra A	Palabra B	DIP	PLS	Global
Cocido	**Crudo**	0.9167	1.0000	1.9167
Secreción-vaginal	**Semen**	0.7667	1.0000	1.7667
Fruta	**Verdura**	0.9275	0.8057	1.7333
Fruto	**Vegetal**	0.9333	0.7156	1.6489
Cuatro	**Tres**	0.5000	0.9447	1.4447
Semana	**Mes**	0.3750	0.9941	1.3691
Blando	**Paladar**	0.9583	0.3720	1.3304
cinco-años	**Superviviente**	0.8333	0.3175	1.1508
Menor	**Mayor**	0.1008	1.0000	1.1008

Tabla B.2: Pares detectados como de *oposición*, para *RCL disyuntivas*.

Palabra A	Palabra B	N. A	N. B	N. Comunes	%Cont	%Dif
Receptor	**Antagónico**	3172	1290	942	73.02	68.511
Bradicardia	**Hipotensión**	425	2521	404	61.41	76.301

Tabla B.3: Pares detectados como de *amplitud*, para el grupo **Diferencias grandes**.

Palabra A	Palabra B	N. A	N. B	N. Comunes	%Cont	%Dif
Diarrea	**Dolor-abdominal**	2824	1877	1557	69.37	34.47
Fiebre	**Tos**	2931	1925	1309	68.00	36.62
Demencia	**Enfermed-alzheimer**	2063	346	266	65.32	62.50
Percepción	**Memoria**	1509	3200	1227	65.28	61.56
Memoria	**Lenguaje**	3024	2041	1321	64.72	35.78
Motor	**Sensitivo**	3109	1318	1147	64.57	65.20
Volumen	**Minuto**	2925	1503	1091	63.87	51.77
Cerdo	**Cisticercos**	1655	2708	1293	61.45	38.33
Célula	**Crecimiento**	3185	1865	1244	60.05	48.05
Queso	**Leche**	1296	2380	1003	59.10	39.46
Suplemento	**Calcio**	1470	2657	851	57.89	43.21
Habla	**Lenguaje**	1171	2217	661	56.45	38.08
Hormona	**Insulina**	2539	1407	763	54.23	41.21
Alergia	**Asma**	2452	1416	1014	53.53	37.71
Candida	**Especie**	1027	2050	520	50.63	37.24
Semen	**Flujo**	1355	3133	937	48.41	64.73
Jeringa	**Aguja**	1181	2449	764	45.89	46.16
Sífilis	**Secundaria**	1837	2984	958	45.02	41.75
Dolor	**Fiebre**	1324	3083	727	44.34	64.03
Contaminada	**Jeringa**	2443	1181	720	44.12	45.94
Intenso	**Duración**	1142	2304	582	43.96	42.30
Tercero	**Cuarto**	2147	1159	491	42.36	35.97
Enfermedad-mental	**Emocional**	639	2062	269	42.10	51.80
Antebrazo	**Mano**	956	2471	400	41.84	55.15
Cadera	**Muñeca**	2410	925	724	41.84	54.06
Genes	**Cromosomas**	2207	830	523	39.88	50.13
Sensibilidad	**Especifico**	2501	1350	533	39.48	41.90
Niño	**Adulto**	826	1988	301	35.23	42.30

Tabla B.4: Pares detectados como de *amplitud*, para el grupo **Diferencias medianas**.

Palabra A	Palabra B	DPS	N. A	N. B	N. Comunes	%Cont	%Dif
abdominal	**diarrea**	2.30	1908	2648	1069	56.03	26.94
abdomen	**torax**	1.17	2297	1826	1109	60.73	17.15
agónico	**antagónico**	3.31	1381	1510	997	60.97	4.70
alimento	**agua**	1.82	2775	2656	1427	44.13	4.33
ampolla	**llaga**	1.20	1566	2020	1119	71.46	16.53
anorexia	**palidez**	2.25	2387	1714	1373	61.44	24.50
ano	**pene**	4.00	1734	1530	1067	59.28	7.43
ano	**recto**	1.29	1734	1964	961	40.25	8.37
ansiedad	**depresión**	1.20	2553	1841	1135	61.65	25.92
aprendizaje	**escritura**	2.35	2152	1649	1001	60.70	18.31
ataque	**derrame-cerebral**	2.82	1782	1552	748	48.20	8.37
auditivo	**visual**	0.85	1473	2217	992	51.80	27.08
axila	**cuello**	0.70	2141	2504	1163	54.32	13.21
blanco	**negro**	1.33	1919	1468	864	50.95	16.42
carne	**animal**	1.56	1881	2167	933	49.60	10.41
carne	**cisticerco**	2.68	2017	2708	1517	60.93	25.15
carne	**crudo**	3.30	2017	1138	928	62.04	32.00
cereal	**pan**	0.56	1207	728	588	75.14	17.44
cerebro	**médula-espinal**	1.65	2445	2135	919	43.04	11.29
Cocido	**carne**	1.92	1528	2017	1373	80.24	17.80
colon	**recto**	0.82	2504	1964	1474	67.41	19.66
columna	**cadera**	1.35	2553	2410	1793	61.78	5.21
confusión	**letra**	3.75	1765	1692	1098	55.97	2.66
contaminada	**aguja**	1.21	2266	2207	1026	46.49	2.15
corto	**acción**	1.12	2559	2015	1045	44.96	19.80
cuello	**cabeza**	1.03	2617	2530	1653	65.34	3.17
dedo	**pie**	2.20	2155	2444	1186	55.03	10.52
dieta	**ejercicio**	2.00	1907	2224	956	50.13	11.54
dolor-cabeza	**fiebre**	4.80	2114	2923	1459	69.02	29.45
duodeno	**porción**	1.13	1278	2024	966	58.69	27.16
escritura	**lectura**	0.78	1835	1766	1194	60.08	2.51
estreñimiento	**diarrea**	2.09	2273	2552	1257	55.30	10.16
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Palabra A	Palabra B	DPS	N. A	N. B	N. Comunes	%Cont	%Dif
estrés	**ansiedad**	0.57	2612	2493	1408	56.48	4.33
fiebre	**malestar**	1.67	3083	2725	2049	64.59	13.03
flexión	**rotación**	3.00	1731	2072	1352	63.55	12.41
fractura	**luxación**	1.59	3031	2489	1640	54.16	19.73
fruta	**cereal**	3.25	1989	1207	888	54.52	28.47
garganta	**boca**	0.69	1763	2547	1216	68.97	28.54
garganta	**nariz**	1.25	1743	1983	877	50.32	8.74
genital	**boca**	2.38	2257	2571	1288	57.07	11.43
glóbulos-rojos	**plaquetas**	0.56	1890	2077	1296	53.17	6.81
grasa	**colesterol**	1.60	2513	2164	1310	60.54	12.70
hemodiálisis	**diálisis**	4.22	1553	1975	1101	57.50	15.36
hombre	**mujer**	2.42	1687	2230	754	39.00	19.77
hongo	**bacteria**	1.87	2206	1480	706	47.70	26.43
humano	**animal**	2.03	2369	2202	1263	57.36	6.08
interno	**externo**	2.50	2878	2250	1516	67.38	22.86
labio	**boca**	1.09	2010	2718	1527	57.41	25.77
larva	**huevo**	2.69	1926	2510	1693	74.04	21.26
leche	**derivado**	0.67	2254	2133	785	36.80	4.40
lenguaje	**escritura**	1.46	2173	1426	950	66.62	27.19
letra	**lectura**	4.71	1692	1766	1064	50.47	2.69
lumbar	**cervical**	1.00	2163	2630	1473	68.10	17.00
lumbar	**dorsal**	0.50	2361	2573	1798	57.73	7.72
masculino	**femenino**	1.81	935	1731	572	52.94	28.98
minuto	**hora**	6.63	1503	1254	552	36.92	9.06
nasal	**tos**	2.18	2668	1826	1099	60.19	30.65
ojos	**boca**	1.27	2545	2510	1366	54.42	1.27
palabra	**letra**	3.06	1544	1692	846	43.72	5.39
pequeño	**grande**	3.39	1561	1744	838	50.29	6.66
picor	**sensación**	1.78	1651	2444	851	51.54	28.87
pierna	**brazo**	0.57	2015	1897	1403	65.58	4.30
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Palabra A	Palabra B	DPS	N. A	N. B	N. Comunes	%Cont	%Dif
pierna	**pie**	3.20	2015	2654	1303	54.69	23.26
positivo	**negativo**	1.53	2417	2379	1135	43.80	1.38
posterior	**anterior**	2.24	2231	1935	1369	67.08	10.78
potasio	**sodio**	0.73	941	1546	763	70.14	22.02
progreso	**lento**	0.75	1879	1283	726	52.84	21.70
proximal	**distal**	1.10	1924	2032	1222	49.06	3.93
pulmón	**corazón**	2.65	2624	2561	1399	54.63	2.29
radioterapia	**quimioterapia**	1.34	1152	1826	494	42.88	24.54
recurrente	**herpes**	1.59	1158	1959	609	52.59	29.16
relacion-sexual	**anal**	1.40	2314	1994	1055	52.91	11.65
rural	**urbano**	4.44	1415	966	709	66.56	16.35
sal	**agua**	1.73	2487	2656	1317	45.03	6.15
secundaria	**primaria**	0.75	2881	2157	1331	61.71	26.36
sentido	**nariz**	3.44	2730	1912	1477	77.25	29.78
sífilis	**gonorrea**	0.92	1706	1958	1002	58.73	9.17
superior	**inferior**	2.62	2526	1958	1094	55.87	20.68
tabaco	**alcohol**	0.67	2169	1939	1236	57.92	8.37
tabaco	**droga**	1.39	2169	1843	970	46.01	11.87
tejido	**hueso**	3.02	2305	3113	1287	48.50	29.41
tensión	**ansiedad**	2.23	2633	2499	1501	60.06	4.88
tensión	**cardiaco**	2.45	2636	2510	1468	58.49	4.59
tibia	**femur**	0.88	1063	1364	635	49.20	10.96
trastorno	**depresión**	4.12	2149	2052	1012	41.47	3.53
vagina	**ano**	2.64	1879	1734	1322	66.96	5.28
vagina	**recto**	0.74	1879	1530	1146	66.27	12.70
vaginina	**pene**	1.22	1750	1755	679	38.80	0.18
vejiga	**recto**	1.56	1938	1964	1024	41.23	0.95
vitamina	**calcio**	3.04	2296	2624	1231	53.61	11.94
vitamina	**mineral**	0.85	2363	2417	1150	48.67	1.97

Tabla B.5: Pares detectados como de *amplitud*, para el grupo **Diferencias pequeñas**.

Apéndice C

Resultados de la Evaluación

Palabra A	Palabra B	J1	J2	J3	J4	$\overline{x}$	Ac_J
oral	**anal**	0.75	1	0.75	1	0.875	0.333
estrés	**ansiedad**	0.75	0.75	0.75	0.75	0.75	1
secreción-vaginal	**semen**	1	0	0.75	0	0.4375	0.333
fruta	**verdura**	0.75	0.75	0.5	0.75	0.6875	0.666
fruto	**vegetal**	0.75	0.75	0.5	0.75	0.6875	0.666
blando	**paladar**	0.75	0.75	0.75	0.5	0.6875	0.666
						$G_p \cong 0.687$	$G_a \cong 0.61$

Tabla C.1: Evaluación de los resultados del sistema, identificados como de *oposición*.

Palabra A	**Palabra B**	**J1**	**J2**	**J3**	**J4**	$\overline{x}$	Ac_J
tensión	**cardiaco**	0	0	1	0.5	0.375	0.333
genital	**voca**	1	0	0	0	0.25	0.666
nasal	**tos**	0	0	1	0	0.25	0.666
recurrente	**herpes**	0	1	0	0.5	0.375	0.333
vejiga	**recto**	1	0	0	0.5	0.375	0.333
vagina	**recto**	0.75	0	0	0.5	0.3125	0.333
alergia	**asma**	0.75	1	0.75	0.5	0.75	0.333
estreñimiento	**diarrea**	0.75	0.75	0.75	0.75	0.75	1
fractura	**luxación**	0.75	0.75	1	0.5	0.75	0.333
bradicardia	**hipotensión**	1	1	1	1	1	1
hemodiálisis	**diálisis**	1	1	1	1	1	1
trastorno	**depresión**	1	1	0.75	1	0.9375	0.666
vitamina	**calcio**	0.75	1	1	1	0.9375	0.666
flexión	**rotación**	1	1	0.75	1	0.9375	0.666
ansiedad	**nervios**	1	1	0.75	1	0.9375	0.666
abdominal	**diarrea**	1	0	1	0.5	0.625	0.333
anorexia	**palidez**	1	1	0.5	1	0.875	0.666
tensión	**ansiedad**	1	0.75	1	0	0.6875	0.333
hongo	**vacteria**	0.75	0	1	1	0.6875	0.333
(glóbulos-rojos)	**plaquetas**	1	1	0.75	1	0.9375	0.666
lumbar	**dorsal**	1	1	1	0.5	0.875	0.666
diarrea	**dolor-abdominal**	1	1	1	0.5	0.875	0.666
fiebre	**tos**	1	1	1	1	1	1
demencia	**(enfermedad-Alzheimer)**	1	1	1	1	1	1
suplemento	**calcio**	0	1	1	1	0.75	0.666
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Palabra A	**Palabra B**	**J1**	**J2**	**J3**	**J4**	$\overline{x}$	Ac_J
hormona	**insulina**	0.75	1	0.5	1	0.8125	0.333
dolor	**fiebre**	1	1	0.75	1	0.9375	0.666
genes	**cromosomas**	1	1	1	1	1	1
(dolor-de-cabeza)	**fiebre**	1	1	1	1	1	1
picor	**sensación**	1	1	1	1	1	1
cerebro	**médula-espinal**	1	1	1	1	1	1
grasa	**colesterol**	1	1	1	1	1	1
(relación-sexual)	**anal**	0	1	0.5	1	0.625	0.333
columna	**cadera**	1	1	1	1	1	1
radioterapia	**quimioterapia**	0.75	1	1	1	0.9375	0.666
garganta	**nariz**	1	1	1	0.5	0.875	0.666
ansiedad	**depresión**	1	1	1	1	1	1
abdomen	**torax**	1	1	1	1	1	1
labio	**boca**	1	1	1	1	1	1
cuello	**cabeza**	1	1	1	1	1	1
lumbar	**servical**	1	1	1	1	1	1
sífilis	**gonorrea**	1	1	0.75	1	0.9375	0.666
vitamina	**mineral**	0.75	1	1	1	0.9375	0.666
colon	**recto**	1	1	1	1	1	1
ataque	**(derrame-cerebral)**	0	1	1	0	0.5	0.333
axila	**cuello**	1	0	0	1	0.5	0.333
						$G_p \cong 0.81$	$G_a \cong 0.7$

Tabla C.2: Evaluación de los resultados del sistema, identificados como de *amplitud*.

Palabra A	Palabra B	J1	J2	J3	J4	$\overline{x}$	Ac_J
VIH	**infección**	0.5	0.5	0.5	0.5	0.5	1
vertebral	**cadera**	0.5	0.5	0.5	0.5	0.5	1
vejiga	**orina**	0.5	0.5	0.5	1	0.625	0.666
síntesis	**ADN**	0.5	0.25	0.5	1	0.5625	0.333
sífilis	**primaria**	0.5	0.5	0.5	0.5	0.5	1
sífilis	**chancros**	0.5	0.5	0.5	0.5	0.5	1
nasal	**sinusitis**	0.5	0.5	0.5	1	0.625	0.666
médula-ósea	**quimioterapia**	0.5	0.25	0.5	0.25	0.375	0.333
lumbar	**punción**	0.5	0.25	0.5	1	0.5625	0.333
llaga	**genitales**	0.25	0.25	0.25	1	0.4375	0.666
linfocitos	**infiltrar**	0.25	0.25	0.25	0.25	0.25	1
infusión	**planta**	0.25	0.25	0.5	0.25	0.3125	0.666
infusión	**agua**	0.5	0.5	0.5	1	0.625	0.333
infarto	**corazón**	0.5	0.5	0.5	1	0.625	0.666
gestación	**embarazo**	0.5	0.5	0.5	0.5	0.5	0.666
genes	**mutación**	0.5	0.5	0.5	1	0.625	0.666
fractura	**osteoporosis**	0.5	0.5	0.5	1	0.625	0.666
espina	**apófisis**	0.5	0.5	0.5	1	0.625	0.333
eritrocitos	**membrana**	0.5	0.25	0.5	0.5	0.4375	0.666
dosis	**efecto**	1	0.5	0.5	1	0.75	0.333
diagnóstico	**tratamiento**	0.5	0.5	0.5	1	0.625	0.666
diabetes	**insulina**	0.5	0.5	0.5	1	0.625	0.333
bilis	**conducta**	0.25	0.25	0.25	0.25	0.25	1
aspirar	**punción**	1	0.5	0.5	0.5	0.625	0.333
asfixia	**hipoxia**	0.5	0.5	0.5	0.5	0.5	1
aprendizaje	**dislexia**	0.5	0.5	1	0.5	0.625	0.333
apnea	**sueño**	0.5	0.5	0.25	1	0.5625	0.333
anticuerpo	**eritrocitos**	0.5	0.25	1	0.25	0.5	0.333
anticuerpo	**inmune**	0.5	0.5	0.5	0.5	0.5	0.666
						$G_p \cong 0.53$	$G_a \cong 0.62$

Tabla C.3: Evaluación de los resultados no identificados por el sistema.

Apéndice D

ER para relaciones de *oposición*

Nr	Expresión Regular	Peso
1	word* [**de** | **del** | **desde**]+ word* Ant [**a** | **al** | **hasta**]+ word* Ant word*	12
2	word* [**de** | **del** | **desde**]+ word* Ant **o** word* Ant word*	7
3	word* Ant word{0,4} [**y**|**o**] word{0,4} Ant word*	2
4	word* [**la** | **las** | **el** | **los** | **sus** | **mas** | **menos** | **son** | **en**]+ Ant **o** word* Ant word*	6
5	word* Ant, **pero** word* Ant word*	16
6	word* Ant [word* |,|;|:] {0,3} [,|;|:]+ [**sino** | **en cambio** | **sin embargo**]+ word{0,3} Ant word*	10
7	word* **tanto** [word* |,|;|:] {0,3} Ant [word* | , | ; | :]{0,3} **como** [word* | , | ; | :]{0,3} Ant word*	9
8	word* [**ambos** | **ambas** | **todo** | **todos** | **toda** | **todas** | **dos** | **entre**]+ [word* | , | ; | :]{0,3}: [word* | , | ; | :]{0,3} Ant **y** [word* | , | ; | :]{0,3} Ant word*	10
9	word* Ant **y** Ant word*	6
10	word* Ant word{0,2}, word{0,3} Ant word*	0.5
⋮	⋮	⋮

Nr	Expresión Regular	Peso
11	word* [**ni** | **no**]+ [word* | , | ; | :]{0,3} Ant [word* | , | ; | :] {0,3} [**ni** | **no** | **tampoco**]+ [word* | , | ; | :]{0,3} Ant word*	5
12	word* Ant [word* |,|;|:]{0,2}: [word* | , | ; | :]{0,3} Ant word*	1
13	word* Ant **o** Ant word*	8
14	word* Ant word{0,3} [**pero** | **sino** | **versus** | **contraposicion** | **ya que** | **contra** | **contraponer** | **sin** | **con el**]+ word{0,3} Ant word*	2
15	word* [**de** | **del** | **desde**]+ word{0,1} Ant word{0,2} [**a** | **al** | **hasta**]+ word{0,3} Ant word*	2
16	word* [**mixto** | **combinado** | **entre**]+ word{0,2} Ant **y** Ant word*	3
17	word* [word* |,|;|:]{0,3}: [word* |,|;|:]{0,3} [**distinto** | **diferente** | **desigual**]+ [word* |,|;|:]{0,3} Ant word*	7
18	word*, [**ambos** | **ambas** | **todo** | **toda** | **todas** | **todos** | **dos** | **entre**]+ Ant [word* |,|;|:]{0,3} Ant word*	5
19	word* Ant [**integrado por** | **compone de** | **miembro de** | **pertenece a** | **entre las** | **entre los** | **entre sus** | **dentro de** | **parte de** | **grupo de** | **constituye a** | **coleccion de** | **unidad de** | **sobre el** | **sobre la** | **sobre las**| **sobre su** | **sobre sus**]+ Ant word*	4
20	word* **en** word{0,1} Ant word{0,3} [**y**|**o**|**u**]+ word{0,2} Ant **de** word*	10
21	word* Ant word{0,3} [**es mayor que**|**es menor que**| **es igual que**]+ word{0,4} Ant word*	10
22	word+, word{0,1} Ant word{0,3} [**y**|**o**|**u**]+ word{0,2} Ant word{0,4}, word*	7

Tabla D.1: ER para opuestos y sus pesos.

Apéndice E

ER para relaciones de *amplitud*

Nr	Expresión Regular	Peso
1	word* Amp word{0,4} [**tal como**|**tales como**]+ word{0,18} Amp word{0,5}, word*	12
2	(word* [**tal**|**tales**]+ word{0,2} Amp word{0,2} [**son**|**es**|**como**]+ word{0,5} Amp+ word{0,5}, word*)| (word* [**tal**|**tales**]+ word{0,2} Amp word{0,2} [**son**|**es**|**como**]+ word{0,5}, Amp+ word{0,5}, word*)	20
3	word* Amp [**integrado por** | **compone de** | **miembro de** | **pertenece a** | **entre las** | **entre los** | **entre sus** | **dentro de** | **parte de** | **grupo de** | **constituye a** | **coleccion de** | **unidad de** | **sobre el** | **sobre la** | **sobre las**| **sobre su** | **sobre sus**]+ Amp word*	5
4	word* [word* |,]{0,18} Amp [word* |,]{0,18} [**u otro** | **o otro** | **y otro**]+ word{0,5} Amp word*	5
5	word* [word* |,]{0,18} Amp [word* |,]{0,18}, [**u otro** | **o otro** | **y otro**]+ word{0,5} Amp word*	12
6	(word* Amp [word* |,]{0,18}, **incluyendo** [word* |,]{0,18} Amp word*) | (word* Amp [word* |,]{0,18}, **incluyendo** [word* |,|;|:] {0,10} [**o**|**y**|**u**]+ word{0,3} Amp word*)	12
⋮	⋮	⋮

Nr	Expresión Regular	Peso
7	(word* [word* |,]{0,18} Amp [word* |,]{0,10}, [**especial**|**especialmente**]+ [word* |,]{0,12} Amp [word* |,] {0,18} word*) | (word* [word* |,]{0,10}, [**especialmente**| **especial**]+ word{0,8} [**o**|**y**|**u**]+ word{0,3} Amp word*)	10
8	word* [**otro**|**otra** | **otros**|**otras**]+ word{0,5} Amp word{0,5} **como:** [word* |,]{0,18} Amp word*	6
9	word* Amp word{0,2}, [word* |,]{0,12} Amp word*	2
10	word* Amp [word*|,|;|:]{0,3} **(sobre todo en)**+ [word*|,]{0,6} Amp word*	13
11	word* Amp word{0,3} [**enclavado en**|**radica en** | **dentro de**| **que corresponde a**]+ word{0,4} Amp word*	6
12	word* **en** word{0,1} Amp word{0,3} [**y** | **o**| **u**]+ word{0,2} Amp **de** word*	6
13	word* Amp word{0,4} [**que aproxima al** | **que aproxima a** | **que nos aproximan a**|**que aproximan al** | **que nos aproxima a**]+ word{0,4} Amp word*	20
14	word* Amp word{0,2} [**de cualquier** | **de cada uno de los** | **estrechamente ligado a**|**de cada una de las** | **de este**]+ word{0,2} Amp word*	3
15	word* Amp word{0,3} [**es mayor que** | **es menor que** | **es igual que**]+ word{0,4} Amp word*	10
16	word+, word{0,1} Amp word{0,3} [**y** | **o** | **u**]+ word{0,2} Amp word{0,4}, word+	5
17	word* Amp word{0,3}, [**entre otros** | **entre otras** | **entre ellos** |**entre ellas** | **entre los que** | **entre las que**]+ [word* |,|;|:]{0,7} Amp word*	8
18	word* Amp word{0,2} **tiene** [word* |,|;|:]{0,6} Amp word*	5
⋮	⋮	⋮

Nr	Expresión Regular	Peso
19	word* [**otro**|**otra** | **otros**|**otras**]+ [word* |,|;|:]{0,10} Amp word{0,3}, **a saber:** [word* |,]{0,18} Amp word*	8
20	word* Amp word{0,4} [**tal como**|**tales como** | **asi como**]+ word{0,4} [,|:]+ [word* |,]{0,18} Amp word{0,4}, word*	10
21	word* [**otro** |**otra** |**otros** |**otras**]+ word{0,4} Amp word{0,3}, **como** [word* |,]{0,20} Amp word*	10
22	(word+, [**incluso**|**incluyendo**]+ word{0,8} Amp word{0,2}, [word+ |,]{0,18} Amp word*) | (word+, [**incluso**|**incluyendo**]+ word{0,6} Amp word{0,2} [**de**|**del**|**de las**|**de los**]+ word{0,1} Amp word*)	16
23	(word+, [**incluso**|**incluyendo**]+ word{0,10} Amp word{0,3} **como** [:|;]+ [word* |,]{0,18} Amp word*) | (word+, [**incluyendo**| **incluso**]+ word{0,10} Amp word{0,3} **como** word{0,4} Amp word*)	8
24	word* Amp [**es un**|**es una**|**es su**]+ Amp word*	2
25	(word* [**entre los**|**entre las**]+ word{0,1} Amp word{0,2}, [word* |,]{0,18} Amp word*) | (word* [**entre los**|**entre las**]+ word{0,1} Amp word{0,6} Amp word*)	6
26	word* Amp word{0,2}, [**este**|**esta** | **estos**|**estas**]+ word{0,4} Amp word*	8
27	word* Amp word{0,1} [**y**|**o** | **u**]+ word{0,1} Amp word*	3
28	word+, word{0,1} Amp word{0,1} (,word{0,3}){0,10}, word{0,1} Amp word*	6
29	word* Amp [word* |,]{0,8}, word{0,4} [**bajo el nombre de** | **con el nombre de** | **se nombra** | **se denomina**]+ word{0,2} Amp word*	20
30	word* Amp [word* |,]{0,8}, word{0,4} **sobre todo** word{0,2} Amp word*	10
⋮	⋮	⋮

Nr	Expresión Regular	Peso
31	word* Amp word{0,1} [**a semejanza** | **en similitud** | **parecido** | **similar**]+ [**de** | **del** | **con** | **a** | **al**]+ word{0,5} Amp word*	20
32	word* **entre** word{0,1} Amp word{0,1} [**y** | **o**| **u**]+ word{0,1} Amp word*	4
33	word* **en** word{0,1} Amp word{0,10} **que** [**en** | **para**]+ word{0,1} Amp word*	20
34	word* Amp word{0,6} [**esta** | **este** | **estas** | **estos** | **esa** | **ese** | **aquella** | **aquel**]+ word{0,1} Amp word*	2
35	word* Amp word{0,2} [**de** | **del** | **en** | **el**]+ word{0,2} Amp word*	2
36	word* **escasez de** word{0,1} Amp word{0,1} (, word{0,3}) {0,10}, word{0,1}) Amp word*	20
37	(word* Amp word{0,1} **sustituye** [**a** | **al**]+ word{0,1} Amp word*) | (word* Amp word{0,1} **sustituye** [**a** | **al**]+ word{0,2} (, word{0,3}) {0,10}, word{0,1} Amp word*)	20
38	word* Amp word{0,1}, **sustituye** [**a** | **al**]+ word{0,6}) Amp word*	4
39	word+, **a saber** (, word{0,3}) {0,10}, word{0,1} Amp word{0,1} (, word{0,3}) {0,10}, word{0,8}) Amp word*	20
40	word* Amp word{0,1} [**integrado por** | **se compone de** | **miembro de** | **pertenece a** | **pertenece al** | **entre** | **dentro de** | **parte de** | **grupo** | **constituido por** | **coleccion de** | **unidad de** | **sobre**]+ word{0,2} Amp word*	3
41	word* [**el** | **la** | **los** | **las**]+ Amp word{0,1} **de** [**el** | **la** | **los** | **las**]+ word{0,1} Amp word*	3
42	word* Amp [word* |,|;|:]{0,3}: [word* |,|;|:]{0,3} [**distinto** | **diferente** | **desigual**]+ [word* |,|;|:]{0,4} Amp word*	6
43	word* [**de** | **del** | **en**]+ word* Amp **o** word* Amp word*	6

Tabla E.1: ER para amplios y sus pesos.

Apéndice F

thesaurus de Economía

Palabra	Vecinos cercanos
abril	febrero [0.067]
absoluto	relativo [0.245]
abundancia	escasez [0.155]
acceso	agua [0.18] salud [0.215]
aceite	pescado [0.151] harina [0.181]
acentuar	propaganda [0.238]
acorde	cartera [0.139]
acreedor	deudor [0.251] deuda [0.129]
actividad	producto [0.297] trabajo [0.275] realizar [0.286] foro [0.303] hombre [0.271] human [0.281] empresa [0.305] economía [0.187] propio [0.296]
acto	transformación [0.057] benefactor [0.042]
actor	coalición [0.148]
actual	robo [0.011] mundial [0.256]
acumulación	ahorro [0.223]
acusar	blanco [0.069]
adaptar	innovar [0.157]
adelante	reembolso [0.126]
adición	pecuniario [0.181]
administración	público [0.165]
adoptar	decisión [0.246]
aduana	provincia [0.17] canon [0.076]
aduanero	canon [0.075] unión [0.155]
adulto	funcional [0.151]
advertencia	anuncio [0.142]
advertir	lector [0.111]

Palabra	Vecinos cercanos
África	América-latina [0.122] sur [0.116]
agencia	solidaridad [0.096] viaje [0.101]
agotar	recursos-naturales [0.132]
agricultor	semilla [0.08] cosecha [0.134] median [0.12]
agricultura	industria [0.142]
agropecuario	forestal [0.081] primario [0.122]
agua	potable [0.096] acceso [0.18] corriente [0.139] aire [0.062]
ahorro	acumulación [0.223] inversión [0.197]
aire	agua [0.062] burbuja [0.112]
algodón	azúcar [0.185] hilo [0.098] arroz [0.107] café [0.151]
alimentación	vivienda [0.159]
alimentación	peruano [0.082]
alimentario	soberano [0.124] sanidad [0.125]
almacén	cerebro [0.221]
alta	volátil [0.131]
alternativa	escasez [0.189]
amazonas	dios [0.137] tumba [0.182] madre [0.131]
ambiente	recursos-naturales [0.159]
amenaza	entrada [0.164] potencial [0.166] competidor [0.137]
América-latina	África [0.122]
americano	mercado-bursátil [0.056]
amigo	película [0.111]
amigo	turista [0.09]
analogía	producción-inmaterial [0.102]
anexo	victima [0.059]
animal	vegetal [0.084] planta [0.164] caza [0.108]
anormal	ciencia-normal [0.175]
anterior	mención [0.134] periodo [0.248] capítulo [0.172]
anual	promedio [0.197] millones-de-dólares [0.177] mínimo [0.188] habitante [0.202]
anuncio	febrero [0.037] advertencia [0.142]
año	millones-de-dólares [0.161] paso [0.211] periodo [0.244] estimar [0.215] millon-de-dólares [0.016]

Palabra	Vecinos cercanos
años	treinta [0.058] veinte [0.093] millón [0.184] mil [0.207]
aplicación	principio [0.297]
aporte	metodológico [0.179]
apreciar	grafica [0.112]
aprobar	conferencia [0.085] dictamen [0.164] debate [0.113]
arancel	contingente [0.088]
arancelario	unilateral [0.158] barrera [0.176]
arbitraje	tribunal [0.171]
árbol	madera [0.2]
argentina	sistema-financiero [0.162]
arrendar	arrienda [0.203] colon [0.24]
arrienda	arrendar [0.203] colon [0.179]
arroz	papa [0.125] algodón [0.107]
artes	facultad [0.161] perfección [0.174]
artificial	natural [0.116]
artista	medico [0.106]
asiático	china [0.167] crisis [0.074]
asignar	eficiente [0.224]
atención	centro [0.188]
atracción	turista [0.098]
aumento	producto [0.294] precio [0.286] riqueza [0.274] disminución [0.207] pro [0.002] trabajo [0.258] producción [0.299] ingreso [0.271] efecto [0.31] consumo [0.289]
automóvil	fabricante [0.133]
axiomático	deducción [0.195]
ayuda	voluntario [0.158]
azar	correr [0.156]
azúcar	café [0.231] algodón [0.185]
bajo	costo [0.294] ingreso [0.276]precio [0.26] economía [0.139] país [0.212] producto [0.254] producción [0.263] trabajo [0.22] mayor [0.281] capital [0.244] foro [0.241]
balance	compensatorio [0.095]

Palabra	Vecinos cercanos
balanza	cuenta-corriente [0.109] desequilibrio [0.113] equilibrio [0.101] déficit [0.169]
balanza-comercial	superávit [0.144]
bancario	depósito [0.167]
banco	préstamo [0.256] billete [0.131] deposito [0.183] dinero [0.226] crédito [0.266] fondo [0.237]
banco-central	reservas [0.203] tipo-de-cambio [0.207]
banco-de-la-república	encajar [0.135]
banco-mundial	monetario [0.087]
barato	caro [0.235]
barrera	arancelaria [0.176]
barril	petróleo [0.091]
beber	comer [0.075]
benefactor	ciclo [0.251] estado [0.152] productor [0.272] acto [0.042] cumplir [0.252] transformación [0.234]
beneficiario	victima [0.111]
beneficio	costo [0.322] genero [0.31] pago [0.292] productor [0.251]
billete	banco [0.131] letra [0.124]
blanco	negro [0.066] acusar [0.069]
bloqueo	tercer [0.059]
bolsa	wall-street [0.114] caída [0.136]
brasil	india [0.184] chile [0.134] china [0.133]
brecha	grupo-de-países [0.177]
bruto	neto [0.096] producción-nacional [0.201]
burbuja	aire [0.112] Japón [0.092]
burgués	proletario [0.15]
bursátil	caída [0.127] cotización [0.136] evolución [0.122]
café	azúcar [0.231] algodón [0.151]
caída	bolsa [0.136] bursátil [0.127]
cálculo	matemáticas [0.19]
calor	luz [0.104]
callar	denunciar [0.181]

Palabra	Vecinos cercanos
cámara	compañero [0.081]
cambio	moneda [0.247] valor [0.284] tas [0.156]
cambio-climático	capa [0.233]
camino	hierro [0.052]
campaña	electoral [0.109]
canasta	tasa-marginal-de-sustitución [0.199]
canon	ley [0.192] millones-de-dólares [0.186] petrolero [0.092] aduanero [0.075] impuesto [0.196] congreso [0.158] cifra [0.194] aduana [0.076]
cantidad	determinado [0.252]
capa	cambio-climático [0.233] enorme [0.046] consecuente [0.014]
capacidad	límite [0.272]
capital	producto [0.317] trabajo [0.316] interés [0.259] foro [0.299] vale [0.055] valor [0.295] trabajador [0.004] bajo [0.244] país [0.294] tierra [0.162] partido [0.313] mercado [0.297] mero [0.014] producción [0.318] empleo [0.231]
capitales-extranjeros	llegada [0.097]
capítulo	anterior [0.172]
capricho	gusto [0.183]
captación	colocación [0.209]
característica	mental [0.033]
carbón	gas [0.091] petróleo [0.093] limpio [0.06]
carne	madera [0.103] res [0.046] Uruguay [0.14]
caro	barato [0.235]
carretera	puente [0.154] puerto [0.088]
carta	enviar [0.19] acorde [0.139]
caso	ejemplo [0.298] empresa [0.302] estado [0.307] posible [0.277] todo [0.31] partido [0.319] diferente [0.303]
castigo	premio [0.294] refuerzo [0.162]
casual	coincidencia [0.074]

Palabra	Vecinos cercanos
caudal	pirámide [0.11]
caza	homínido [0.165] flecha [0.235] animal [0.108]
celebrar	conferencia [0.09] julio [0.115]
central	gobierno [0.218] lima [0.15] descentralización [0.2]
centralizado	subasta [0.184]
centro	educación [0.241] atención [0.188]
cerebro	información [0.054] almacén [0.221]
cerrar	puerto [0.071]
certificado	digital [0.136] clave-pública [0.197] registro [0.163] usuario [0.22] firma [0.153]
ceso	valor [0.011]
ciclo	cierre [0.129] benefactor [0.251] cumplir [0.26] etapa [0.216]
ciclo-económico	cumplir [0.256]
cien	gasolina [0.076] mil [0.088] progreso [0.056] kilómetro [0.085]
ciencia	estudio [0.282] social [0.189] explicación [0.267] científico [0.288] conocimiento [0.237] investigador [0.266]
ciencia-normal	paradigma [0.182] comunidad-científica [0.214] anormal [0.175]
científico	teoría [0.226] ciencia [0.288] investigador [0.291] método [0.284]
cientos	miles [0.157] veinte [0.169]
cierre	ciclo [0.129]
cierto	manera [0.317] idea [0.24]
cifra	canon [0.194] millones-de-dólares [0.183] representar [0.201] millón [0.2]
cinco	seis [0.161] treinta [0.126] cuatro [0.136]
cinco-años	niño [0.097]
circular	trabajador [0.039]
civil	pacto [0.088]
clásico	visión [0.21] neoclásico [0.148] escuela [0.182]
clasifican	factores-de-producción [0.211]
clave	clave-pública [0.122] criptografía [0.12]
clave-privada	mensaje [0.189] clave-pública [0.277] digital [0.138]
clave-pública	clave [0.122] clave-privada [0.277] certificado [0.197] registro [0.094]
cliente	proveedor [0.1] global [0.12]

Palabra	Vecinos cercanos
coalición	actor [0.148]
cobre	plata [0.168] pieza [0.148]
cohesión	fondo-estructural [0.1]
coincidencia	casual [0.074]
colector	elección [0.216] individual [0.163] decisión [0.181]
colocación	captación [0.209]
colon	arrienda [0.179] arrendar [0.24]
colonia	colonial [0.176]
colonial	colonia [0.176]
comer	beber [0.075]
comerciante	fabricante [0.137]
comercio	libertad [0.2] liberal [0.039] ventaja [0.236] exterior [0.162] industria [0.261]
cometer	error [0.093]
comisión	dictamen [0.117] presidente [0.142] congreso [0.142] europeo [0.132]
compañero	cámara [0.081]
compensación	fondo [0.133] provincia [0.165] municipal [0.119]
compensatorio	balance [0.095]
competente	perfecto [0.218] monopolio [0.241]
competidor	amenaza [0.137]
complementario	concurrente [0.072] sustitución [0.089]
complicación	fáctico [0.057]
comportamiento	individual [0.245] explicación [0.253] racional [0.25] individuo [0.235]
comprador	vendedor [0.307]
comprar	mercancía [0.288] vender [0.287] moneda [0.268]
comunidad-científica	ciencia-normal [0.214] paradigma [0.132]
comunismo	sentido [0.285]
concepción	filosofía [0.216]
concepto	definición [0.261] definir [0.265]
concurrente	complementario [0.072]
condición	socio [0.025]
confederación	sindicato [0.195]
conferencia	aprobar [0.085] celebrar [0.09]

Palabra	Vecinos cercanos
congreso	dictamen [0.107] ley [0.098] comisión [0.142] presidente [0.177] pleno [0.125] parlamentario [0.073] debate [0.178] Luis [0.1] canon [0.158] proyecto [0.151]
conjunto	constituir [0.278]
conocimiento	nuevo [0.297] desarrollo [0.286] ciencia [0.237] tecnología [0.252] información [0.282] llamado [0.266]
consecuente	capa [0.014]
consejo	presidente [0.131] ministro [0.164] transitorio [0.07]
consideración	teoría [0.285] ocio [0.006] economía [0.202] partido [0.319] idea [0.198] fe [0.003]
constitucional	reforma [0.128]
constituir	conjunto [0.278]
construcción	vivienda [0.15]
consulta	página [0.118]
consumidor	productor [0.29]
consumo	riqueza [0.291] producción [0.29] aumento [0.289] objeto [0.286]
contacto	eje [0.068]
contexto	descentralización [0.19]
contingente	arancel [0.088]
controversia	torno [0.077]
convenio	sindical [0.199]
convergente	divergente [0.151] tasa-de-crecimiento [0.121]
corrección	falla [0.132]
correcto	hijo [0.121]
correr	azar [0.156]
corriente	pensamiento [0.168] agua [0.139]
corto	plazo [0.284] largo [0.212] flujo [0.203]
cosa	precio [0.201] naturaleza [0.267]
cosecha	siembra [0.131] agricultor [0.134] superficie [0.077]
costo	beneficio [0.322] producción [0.288] precio [0.274] pago [0.288] bajo [0.294]
costo-marginal	monopolio [0.123]

Palabra	**Vecinos cercanos**
costo-social-de-producción	precio-de-venta [0.23]
cotizar	bursátil [0.136]
creciente	tendencia [0.217]
crecimiento	inflación [0.206]
crecimiento	ritmo [0.093]
crédito	préstamo [0.267] banco [0.266]
criptografía	secreto [0.115] clave [0.12]
crisis	asiática [0.074]
crisis-financiera	ocurrente [0.126]
cruz	subsidio [0.082]
cuadro	siguiente [0.113]
cualitativo	cuantitativo [0.237]
cuantitativo	cualitativo [0.237]
cuarto	trimestre [0.104] escalón [0.101] quinto [0.087]
cuatro	cinco [0.136]
cuenta-corriente	déficit [0.091] balanza [0.109]
cultivo	tierra [0.143] terreno [0.185]
cultural	identidad [0.185]
cumplir	ciclo [0.26] benefactor [0.252] función [0.213] obligación [0.241] ciclo-económico [0.256]
curso	forzosos [0.091]
curva	gráfica [0.148]
chile	brasil [0.134]
china	india [0.178] asiático [0.167] brasil [0.133]
dado	momento [0.289]
daño	reparar [0.139]
dato	estadística [0.162]
debate	congreso [0.178] aprobar [0.113]
débil	fuerte [0.191]
década	siglo [0.211]

Palabra	Vecinos cercanos
decisión	instinto [0.129] adoptar [0.246] individuo [0.216] tomar [0.216] racional [0.24] colector [0.181]
declarar	pacto [0.169]
decreciente	producto-marginal [0.181]
decreto	supremo [0.165]
deducción	deductivo [0.169] axiomático [0.195]
deductivo	inductivo [0.163] deducción [0.169]
defensor	escritor [0.121] libre-comercio [0.117]
déficit	cuenta-corriente [0.091] financiar [0.193] balanza [0.169]
definición	concepto [0.261]
definir	concepto [0.265]
deflación	salvaje [0.118]
dejar	patrimonio [0.079]
demanda	oferta [0.335] precio [0.253] equilibrio [0.254] efecto [0.268] igual [0.265]
democracia	liberal [0.144]
demostrar	experimento [0.227]
denunciar	callar [0.181] mutuo [0.107]
departamento	distrito [0.128] dios [0.117]
dependiente	independiente [0.109]
deposito	banco [0.183] bancario [0.167]
depresión	treinta [0.141] vuelta [0.137]
derecho	internacional [0.254] instrumento [0.226] social [0.226] humano [0.262] información [0.237] laboral [0.186] interno [0.154] nacer [0.049]
derecho-propiedad	intelectual [0.156]
desarrollo	nuevo [0.324] social [0.318] economía [0.226] país [0.297] político [0.311] mayor [0.313] mercado [0.308] tecnología [0.203] nación [0.205] conocimiento [0.286] foro [0.007]
desastre	desplazar [0.134]
descentralización	central [0.2] recaudación [0.17] contexto [0.19]
desempleo	inflación [0.175] desempleado [0.072]
desempleado	desempleo [0.072]

Palabra	Vecinos cercanos
desequilibrio	balanza [0.113]
desplazar	esfuerzo [0.13] desastre [0.134]
determinado	oferta [0.242] cantidad [0.252]
deuda	financiar [0.249] pago [0.224] gasto [0.243] gobierno [0.263] obligación [0.239] acreedor [0.129]
deudor	acreedor [0.251]
devaluación	tipo-de-cambio [0.235]
dictamen	congreso [0.107] comisión [0.117] Luis [0.211] aprobar [0.164]
dicha	manera [0.264]
diez	ocho [0.093] veinte [0.172]
diferencial	ecuación [0.122]
diferente	foro [0.304] país [0.252] caso [0.303]
digital	mensaje [0.162] clave-privada [0.138] firma [0.188] certificado [0.136] nueva-economía [0.126]
dinero	préstamo [0.183] banco [0.226] foro [0.264] moneda [0.264] pago [0.297] utilizar [0.281] persona [0.274] valor [0.263]
dinero-electrónico	tarjeta [0.165]
dios	departamento [0.117] madre [0.254] amazonas [0.137]
directo	indirecto [0.154]
director	instituto [0.077]
discriminación	notorio [0.078]
disminución	aumento [0.207]
disposición	legal [0.151]
distancia	metro [0.13]
distrito	jurisdicción [0.124] transferencia [0.15] departamento [0.128] provincia [0.223]
divergente	convergente [0.151]
divisas	reservas [0.181]
división-del-trabajo	extensión [0.181] especialización [0.138]
doce	mes [0.067]
documento	firma [0.165]

Palabra	Vecinos cercanos
dólar	estados-unidos [0.151] peso [0.143] tipo-de-cambio [0.202]
dotación	inicial [0.086] repartición [0.147]
droga	tráfico [0.156]
duro	golpe [0.112]
ecología	popular [0.116]
economía	trabajo [0.262] teoría [0.163] rol [0.014] foro [0.25] país [0.281] social [0.236] desarrollo [0.226] consideración [0.202] genero [0.166] actividad [0.187] idea [0.079] era [0.072] bajo [0.139] político [0.217] mero [0.006] mercado [0.227] partido [0.259] general [0.193]
economía-bienestar	equilibrio-competitivo [0.131]
economía-instintiva	economía-tradicional [0.193]
economía-social	solidariad [0.168]
economía-tradicional	economía-instintiva [0.193]
ecuación	diferencial [0.122]
edad	niño [0.15]
educación	salud [0.223] superior [0.225] centro [0.241] vivienda [0.167]
efecto	demanda [0.268] aumento [0.31]
eficiente	asignar [0.224]
eje	contacto [0.068]
ejecución	ejecutar [0.041] encargo [0.052]
ejecutivo	legislativo [0.156] ejecutar [0.041] presidente [0.153]
ejemplo	caso [0.298]
elección	colector [0.216]
electoral	campaña [0.109]
eléctrica	gas [0.117] instalar [0.127] energía [0.12]
elemental	lógica [0.112]
empírico	teórico [0.219] hipótesis [0.223]
empleo	trabajo [0.216] capital [0.231]
empresa	trabajo [0.268] persona [0.27] producción [0.293] producto [0.29] mercado [0.301] caso [0.302] actividad [0.305]
empresa-multinacional	frente [0.183]
empresario	obrero [0.167]

Palabra	Vecinos cercanos
encajar	banco-de-la-república [0.135] institución-financiera [0.17]
encargo	ejecución [0.052]
energía	tiempo [0.143] eléctrica [0.12] mina [0.178]
enero	junio [0.071] fecha [0.062] interés [0.006]
enorme	capa [0.046]
entorno	universo [0.185]
entrada	salida [0.158] amenaza [0.164]
enviar	carta [0.19] mensaje [0.075]
epistemología	metafísica [0.175]
equidad	justicia [0.175]
equilibrio	solución [0.236] demanda [0.254] existencia [0.258] oferta [0.275] general [0.199] modelo [0.258] balanza [0.101]
equilibrio-competitivo	óptimo-pareto [0.242] economía-bienestar [0.131]
equipo	maquinaria [0.189]
equitativo	justo [0.135]
era	trabajador [0.016] economía [0.072]
error	cometer [0.093]
escala	mundial [0.214]
escalón	pirámide [0.144] cuarto [0.101]
escape	maestro [0.06]
escasez	alternativa [0.189] abundancia [0.155]
escenario	referente [0.137]
esclavo	muerto [0.17]
escritor	defensor [0.121]
escuela	pensamiento [0.18] neoclásico [0.136] metodología [0.152] clásico [0.182] evolución [0.091]
espacio	gestión [0.183]
especialización	división-del-trabajo [0.138]
establecer	trato [0.294] principio [0.29]

Palabra	Vecinos cercanos
estadística	dato [0.162]
estado	ingreso [0.255] público [0.28] social [0.303] benefactor [0.152] país [0.279] partido [0.312] político [0.304] caso [0.307] todo [0.314]
estados-unidos	Japón [0.145] dólar [0.151]
estatal	intervención [0.196]
estimar	año [0.215] millones-de-dólares [0.173]
estrategia	juego [0.202] global [0.211] jugador [0.17] palanca [0.055]
estrecho	liga [0.111]
estructural	reforma [0.156]
estudio	ciencia [0.282] objeto [0.272] ocio [0.015]
etapa	reposición [0.214] ciclo [0.216]
euro	zona [0.088] europeo [0.077]
europeo	comisión [0.132] euro [0.077]
evasión	recaudación [0.14]
evolución	bursátil [0.122] escuela [0.091]
ex	único [0.01] jefe [0.185]
excepción	regla [0.136]
existencia	equilibrio [0.258]
experiencia	prueba [0.196] demostrar [0.227]
experto	reunión [0.17]
explicación	fenómeno [0.255] teoría [0.254] intentar [0.195] modelo [0.28] ciencia [0.267] comportamiento [0.253]
explícito	implícito [0.182]
explotación	recursos-naturales [0.148]
exportación	importación [0.285]
expresar	término [0.171]
extensión	división-trabajo [0.181]
exterior	comercio [0.162]
externo	interno [0.244]
extracción	reposición [0.212] petróleo [0.126] pesquero [0.102]
extranjero	moneda [0.256] inversión [0.252] nacional [0.248] importación [0.255]
fabricante	automóvil [0.133] comerciante [0.137]

Palabra	Vecinos cercanos
fáctico	complicación [0.057]
factores-de-producción	clasifican [0.211]
facultad	artes [0.161]
falso	verdadero [0.117]
falla	corrección [0.132]
familia	res [0.022]
fe	consideración [0.003] manifestar [0.064] fin [0.004]
febrero	abril [0.067] anuncio [0.037] marzo [0.088]
fecha	enero [0.062]
federal	república [0.142]
fenómeno	observar [0.244] explicación [0.255]
fijación	incidente [0.091]
fijo	tipo-de-cambio [0.207]
filial	matriz [0.216] subsidiario [0.11]
filosofía	concepción [0.216]
fin	medio [0.288] fe [0.004]
financiar	deuda [0.249] gasto [0.217] déficit [0.193]
finito	infinito [0.097]
firma	mensaje [0.104] documento [0.165] digital [0.188] certificado [0.153]
fiscal	monetario [0.221] político [0.097]
flecha	caza [0.235] logística [0.165]
flor	química [0.083]
flujo	corto [0.203] plazo [0.205]
fondo	internacional [0.231] banco [0.237] monetario [0.223] compensación [0.133]
fondo-estructural	reglamento [0.082] cohesión [0.1]
forestal	agropecuario [0.081] pesquero [0.112]
formación	profesional [0.213]
fórmula	matemáticas [0.217]
foro	economía [0.25] partido [0.344] diferente [0.304] actividad [0.303] dinero [0.264] human [0.25] capital [0.299] todo [0.319] bajo [0.241]
foro	tema [0.029] negocio [0.031] organización [0.024] desarrollo [0.007] último [0.013]

Palabra	Vecinos cercanos
forzosos	curso [0.091] desplazar [0.13]
francés	ingles [0.173]
Francia	franco [0.11] vino [0.125]
franco	pieza [0.135] Francia [0.11]
frente	empresa-multinacional [0.183] protección [0.217]
fuerte	débil [0.191]
función	cumplir [0.213]
funcional	adulto [0.151]
función-de-demanda	tasa-de-producción [0.147] lineal [0.168]
fundador	padre [0.091]
futuro	presente [0.242]
garantizar	pro [0.017]
gas	petróleo [0.137] liquido [0.13] eléctrica [0.117] carbón [0.091]
gasolina	kilómetro [0.123] cien [0.076] vehículo [0.093]
gasto	ingreso [0.252] financiar [0.217] impuesto [0.267] público [0.253] inversión [0.244] deuda [0.243]
general	teoría [0.284] término [0.252] equilibrio [0.199] economía [0.193]
genero	beneficio [0.31] nuevo [0.296] economía [0.166] riqueza [0.284]
genética	informática [0.164] ingeniería [0.138]
gestión	urbano [0.174] espacio [0.183]
gira	torno [0.086]
global	estrategia [0.211] estratega [0.06] negocio [0.232] mundial [0.243] cliente [0.12]
gobierno	central [0.218] deuda [0.263]
goce	posesión [0.126]
golpe	duro [0.112]
gradual	lento [0.122]
gráfica	muestra [0.161] apreciar [0.112] curva [0.148]
grande	pequeño [0.219]
grupo-de-países	brecha [0.177]
gubernamental	impulsor [0.057]
guerra	paz [0.092]

Palabra	Vecinos cercanos
gusto	capricho [0.183]
habilidad	operario [0.142] talento [0.18]
habitante	ingreso-natural [0.204] mínimo [0.241] anual [0.202] millón [0.183]
habitar	zona [0.135]
harina	pescado [0.261] aceite [0.181] trigo [0.145]
hecho	teoría [0.283]
hierro	camino [0.052]
hijo	padre [0.183] correcto [0.121]
hilo	algodón [0.098]
hipoteca	pasivo [0.068]
hipótesis	teoría [0.193] empírico [0.223] teórico [0.26]
historia	largo [0.243]
historia	referente [0.19]
hombre	natural [0.241] trabajo [0.242] naturaleza [0.248] actividad [0.271] human [0.292] necesidad [0.291] riqueza [0.292] objeto [0.288]
homínido	caza [0.165]
hora	semana [0.108] paño [0.102] jornada [0.173]
human	foro [0.25] hombre [0.292] derecho [0.262] información [0.257] actividad [0.281] objeto [0.286]
humo	patrón-oro [0.187]
idea	economía [0.079] consideración [0.198] partido [0.153] cierto [0.24]
identidad	cultural [0.185]
igual	demanda [0.265] oferta [0.256]
implementación	recomendar [0.136]
implícito	explícito [0.182]
imponer	restricción [0.197]
importación	total [0.252] extranjero [0.255] exportación [0.285]
impuesto	renta [0.279] canon [0.196] ingreso [0.257] pago [0.227] gasto [0.267] recaudación [0.131]
impulsor	gubernamental [0.057]
incidente	fijación [0.091]
independiente	dependiente [0.109]

Palabra	Vecinos cercanos
india	china [0.178] brasil [0.184]
individual	social [0.173] comportamiento [0.245] colector [0.163] trabajador [0.016]
indirecto	directo [0.154]
individuo	sociedad [0.282] decisión [0.216] manera [0.29] comportamiento [0.235]
inductivo	deductivo [0.163]
industria	agricultura [0.142] trabajador [0.017] comercio [0.261]
industrialización	potencia [0.15] reciente [0.124]
inferior	superior [0.155]
infinito	finito [0.097]
inflación	monetario [0.221] desempleo [0.175] crecimiento [0.206] tasa [0.22]
información	cerebro [0.054] derecho [0.237] tecnología [0.238] conocimiento [0.282] tiempo [0.272] human [0.257]
informática	telecomunicaciones [0.178] genética [0.164]
infraestructura	red [0.182]
ingeniería	genética [0.138] telecomunicaciones [0.141]
ingles	francés [0.173]
ingreso	res [0.007] aumento [0.271] gasto [0.252] bajo [0.276] estado [0.255] impuesto [0.257]
ingreso-natural	habitante [0.204]
inicial	dotación [0.086]
innovar	adaptar [0.157]
inseguro	sede [0.054]
instalación	multinacional [0.172] electrónica [0.127] potencia [0.11]
instancia	último [0.074]
instinto	tradición [0.154] racional [0.171] tradicional [0.186] decisión [0.129]
institución-financiera	encaje [0.17]
instituto	director [0.077]
instrumental	realista [0.147]
instrumento	derecho [0.226]
intelectual	derecho-propiedad [0.156]
intentar	explicación [0.195]
interés	pago [0.298] capital [0.259] enero [0.006]

Palabra	**Vecinos cercanos**
internacional	mercado [0.245] nacional [0.279] derecho [0.254] interno [0.164] monetario [0.177] fondo [0.231] órgano [0.152]
interno	mercado [0.099] externo [0.244] internacional [0.164] derecho [0.154]
intervención	estatal [0.196]
inversión	extranjero [0.252] ahorro [0.197] gasto [0.244]
investigador	programa [0.242] científico [0.291] método [0.272] ciencia [0.266]
isla	riqueza-propia [0.188]
Japón	estados-unidos [0.145] tasa-de-interés [0.104] burbuja [0.092]
jefe	ex [0.185]
jornada	horario [0.173]
jornal	obrero [0.144]
juego	jugador [0.173] estrategia [0.202] papel [0.212] teórico [0.221]
jugador	juego [0.173] estrategia [0.17]
juicio-de-valor	normativa [0.189]
julio	celebrar [0.115]
junio	semana [0.091] enero [0.071]
jurisdicción	distrito [0.124]
justicia	equidad [0.175]
justo	equitativo [0.135]
kilómetro	gasolina [0.123] cien [0.085]
laboral	derecho [0.186] protección [0.234]
largo	corto [0.212] historia [0.243] plazo [0.306]
lavar	prevenir [0.117]
lector	advertir [0.111]
legal	disposición [0.151]
legislativo	ejecutivo [0.156]
lenguaje	usual [0.127]
lento	gradual [0.122]
letra	billete [0.124]
ley	canon [0.192] congreso [0.098]
liberal	democracia [0.144]
liberal	comercio [0.039]

Palabra	Vecinos cercanos
libertad	comercio [0.2]
libre	mercado [0.185]
libre-comercio	defensor [0.117]
liga	estrecho [0.111]
lima	central [0.15] provincia [0.191]
límite	capacidad [0.272]
limpio	carbón [0.06]
línea	tensión [0.094]
lineal	función-de-demanda [0.168]
liquido	gas [0.13]
local	regional [0.202]
lógica	método [0.219] elemental [0.112]
logística	tecnología [0.096] flecha [0.165]
lugar	segundo [0.25] ocupación [0.213] primero [0.265]
Luis	dictamen [0.211] congreso [0.1] presidente [0.114] ministro [0.093]
luz	calor [0.104]
llamado	conocimiento [0.266]
llegada	capitales-extranjeros [0.097]
lleno	vacío [0.044]
madera	árbol [0.2] carne [0.103]
madre	patria [0.108] dios [0.254] amazonas [0.131]
maestro	escape [0.06]
manera	cierto [0.317] dicho [0.264] posible [0.316] individuo [0.29]
manifestar	fe [0.064]
mano	teorema [0.152] precio [0.069]
manufactura	rival [0.125]
manzana	tasa-marginal-de-sustitución [0.217]
maquinaria	equipo [0.189]
marginal	útil [0.158]
marketing	uniforme [0.139] mezcla [0.087]
marzo	febrero [0.088]
matemáticas	modelo [0.179] cálculo [0.19] fórmula [0.217]

Palabra	Vecinos cercanos
matriz	filial [0.216]
mayo	país [0.004]
mayor	menor [0.169] vez [0.31] desarrollo [0.313] bajo [0.281] país [0.302]
mediano	agricultor [0.12]
medico	artista [0.106]
medida	patrón [0.108]
medio	fin [0.288]
mención	anterior [0.134]
menor	mayor [0.169]
mensaje	enviar [0.075] firma [0.104] digital [0.162] clave-privada [0.189]
mental	característica [0.033]
mente	trabajo [0.026]
mercado	internacional [0.245] producto [0.309] economía [0.227] mundial [0.219] desarrollo [0.308] interno [0.099] libre [0.185] país [0.297] capital [0.297] precio [0.284] empresa [0.301]
mercado-bursátil	americano [0.056]
mercancía	precio [0.215] valor [0.214] necesidad [0.254] moneda [0.268] compra [0.288]
merecer	pena [0.094]
mero	economía [0.006] resultado [0.021] país [0.011] capital [0.014]
mes	doce [0.067]
metafísica	epistemológico [0.175]
metal	oro [0.27] moneda [0.131] plata [0.255]
metales-preciosos	oro [0.194] plata [0.187]
método	lógica [0.219] científico [0.284] utilizar [0.213] investigador [0.272]
metodológico	escuela [0.152] aporte [0.179]
metro	distancia [0.13]
mezcla	marketing [0.087]

Palabra	Vecinos cercanos
mil	cientos [0.157] millones-de-dólares [0.176] cien [0.088] millón [0.24] años [0.207] millón-de-dólares [0.021]
millón	habitante [0.183] años [0.184] mil [0.24] cifra [0.2]
millón-de-dólares	mil [0.021] año [0.016]
millones-de-dólares	canon [0.186] mil [0.176] total [0.131] año [0.161] cifra [0.183] anual [0.177] estimación [0.173]
mina	ministerio [0.121] oro [0.158] energía [0.178]
mínimo	habitante [0.241] anual [0.188]
ministerio	mina [0.121]
ministro	Luis [0.093] presidente [0.201] consejo [0.164] público [0.024]
modelo	matemáticas [0.179] teoría [0.258] explicación [0.28] equilibrio [0.258]
modelo-de-competencia-perfecta	subasta [0.142]
momento	dad [0.289]
moneda	mercancía [0.268] metal [0.131] oro [0.154] cambio [0.247] dinero [0.264] extranjero [0.256] compra [0.268]
monetario	internacional [0.177] fondo [0.223] político [0.122] fiscal [0.221] inflación [0.221] banco-mundial [0.087]
monopolio	costo-marginal [0.123] competente [0.241]
monto	recaudación [0.192]
mover	vehículo [0.067]
móvil	teléfono [0.053]
movilizar	símbolo [0.064]
mueble	vestido [0.121]
muerto	esclavo [0.17]
muestra	gráfica [0.161]
multinacional	instalación [0.172] territorio [0.179]
mundial	actual [0.256] global [0.243] mercado [0.219] escala [0.214] nivel [0.259]
mundo	tercer [0.11] vez [0.296]
municipal	provincial [0.177] compensación [0.119]
mutuo	denuncia [0.107] recíproco [0.167]

Palabra	Vecinos cercanos
nacer	niño [0.119] derecho [0.049]
nación	desarrollo [0.205]
nacional	internacional [0.279] extranjero [0.248] nivel [0.275]
natural	hombre [0.241] artificial [0.116]
naturaleza	cosa [0.267] hombre [0.248]
necesario	trabajador [0.008]
necesidad	mercancía [0.254] hombre [0.291] valor [0.27] objeto [0.313] recursos [0.283]
negativo	positivo [0.222]
negocio	global [0.232] foro [0.031]
negro	blanco [0.066]
neoclásico	ortodoxo [0.071] clásico [0.148] escuela [0.136]
neto	bruto [0.096]
new-york	timidez [0.188] stock [0.052] paúl [0.15]
niño	edad [0.15] cinco-años [0.097] nacer [0.119]
nivel	mundial [0.259] nacional [0.275]
nominal	tipo-de-cambio [0.174]
normal	regular [0.221]
normativa	juicio-de-valor [0.189] positivo [0.177] postura [0.141]
norte	sur [0.196]
notorio	discriminación [0.078]
novedad	variedad [0.088]
nueva-economía	vieja [0.099] digital [0.126]
nuevo	desarrollo [0.324] tecnología [0.214] genero [0.296] conocimiento [0.297]
número	reducción [0.257]
objeto	estudio [0.272] teoría [0.272] hombre [0.288] human [0.286] necesidad [0.313] todo [0.301]consumo [0.286]
obligación	deuda [0.239] cumplir [0.241]
obligatorio	providencia [0.115]
obrero	empresario [0.167] jornal [0.144]
observar	fenómeno [0.244]

Palabra	Vecinos cercanos
occidente	país-subdesarrollado [0.1]
ocio	trabajo [0.004] estudio [0.015] consideración [0.006] social [0.005]
ocupación	lugar [0.213]
ocurrencia	probabilidad [0.108] crisis-financiera [0.126]
ocho	diez [0.093]
oferta	determinado [0.242] demanda [0.335] equilibrio [0.275] igual [0.256] precio [0.184]
operario	habilidad [0.142]
óptimo-pareto	equilibrio-competitivo [0.242]
orden	racional [0.221]
organización	persona [0.222] foro [0.024]
órgano	internacional [0.152]
oro	vale [0.189] plata [0.363] moneda [0.154] metal [0.27] mina [0.158] metales-preciosos [0.194]
ortodoxo	neoclásico [0.071]
pacifico	sur [0.063]
pacto	declarar [0.169] civil [0.088]
padre	hijo [0.183] fundador [0.091]
página	consulta [0.118]
pago	costo [0.288] puesto [0.153] beneficio [0.292] deuda [0.224] impuesto [0.227] interés [0.298] precio [0.266] dinero [0.297]
país	producto [0.298] trabajo [0.296] desarrollo [0.297] economía [0.281] trabajador [0.003] estado [0.279] bajo [0.212] político [0.268] mayo [0.004] capital [0.294] todo [0.288] mero [0.011] mayor [0.302] partido [0.316] mercado [0.297] diferente [0.252]
país-en-desarrollo	pobreza [0.232]
país-pobre	país-rico [0.193]
país-rico	país-pobre [0.193]
país-subdesarrollado	occidente [0.1]
palanca	estrategia [0.055]
pan	trigo [0.121]
paño	pieza [0.114] vino [0.259] hora [0.102]

Palabra	Vecinos cercanos
papa	arroz [0.125]
papel	juego [0.212]
paradigma	sustitución [0.173] ciencia-normal [0.182] comunidad-científica [0.132]
parlamentario	congreso [0.073]
partido	trabajo [0.318] foro [0.344] social [0.317] político [0.299] economía [0.259] consideración [0.319] idea [0.153] estado [0.312] pro [0.001] capital [0.313] país [0.316] caso [0.319]
pasado	año [0.211]
pasivo	hipoteca [0.068]
patria	madre [0.108]
patrimonio	dejar [0.079]
patrón	medida [0.108]
patrón-oro	humo [0.187]
paúl	new-york [0.15] timidez [0.146]
paz	guerra [0.092]
pecuniario	adición [0.181]
película	amigo [0.111] premio [0.056]
peligro	socio [0.098]
pena	merecer [0.094]
pensamiento	escuela [0.18] corriente [0.168]
pequeño	grande [0.219]
perfección	artes [0.174]
perfecto	competente [0.218]
periodo	año [0.244] anterior [0.248]
persona	organización [0.222] empresa [0.27] dinero [0.274] recursos [0.283] realizar [0.289]
peruano	promedio [0.171] pesquero [0.098] alimentación [0.082]
pescado	harina [0.261] aceite [0.151]
peso	dólar [0.143]
pesquero	peruano [0.098] forestal [0.112] extracción [0.102]
petróleo	gas [0.137] extracción [0.126] carbón [0.093] barril [0.091]

Palabra	Vecinos cercanos
petrolero	canon [0.092]
pib-per-capita	ritmo [0.088]
pieza	cobre [0.148] paño [0.114] plata [0.149] franco [0.135]
pirámide	escalón [0.144] caudal [0.11]
planta	animal [0.164]
plata	cobre [0.168] pieza [0.149] oro [0.363] metal [0.255] metales-preciosos [0.187]
plazo	corto [0.284] flujo [0.205] largo [0.306]
pleno	congreso [0.125]
población	pobreza [0.227] pobre [0.245]
población-mundial	quinto [0.117]
pobre	rico [0.246] población [0.245]
pobreza	población [0.227] país-en-desarrollo [0.232]
político	público [0.274] social [0.319] desarrollo [0.311] país [0.268] partido [0.299] monetario [0.122] fiscal [0.097] resultado [0.298] problema [0.277] tarea [0.018] sistema [0.29] estado [0.304] tema [0.126] economía [0.217]
ponderado	voto [0.068]
popular	ecología [0.116]
posesión	goce [0.126]
posibilidad	socio [0.041]
posible	caso [0.277] manera [0.316]
positivo	normativa [0.177] negativa [0.222]
postura	normativa [0.141] torno [0.115]
potable	agua [0.096]
potencia	industrialización [0.15] instalar [0.11]
potencial	amenaza [0.166]
precepto	prudencia [0.144]
precio	cosas [0.201] producto [0.306] mercancía [0.215] producción [0.304] demanda [0.253] aumento [0.286] costo [0.274] pago [0.266] oferta [0.184] valor [0.298] bajo [0.26] mano [0.069] mercado [0.284]
precio-de-venta	costo-social-de-producción [0.23]
pregunta	respuesta [0.216]

Palabra	Vecinos cercanos
premio	castigo [0.294] película [0.056] refuerzo [0.214]
presente	futuro [0.242]
presidente	Luis [0.114] ministro [0.201] congreso [0.177] ejecutar [0.153] consejo [0.131] republica [0.118] comisión [0.142]
préstamo	préstamo [0.228] solicitar [0.168]
préstamo	dinero [0.183] banco [0.256] servicio [0.225] crédito [0.267] préstamo [0.228]
prevención	lavado [0.117]
primario	secundario [0.139] agropecuario [0.122] superávit [0.194]
primero	lugar [0.265] segundo [0.269]
primero	segundo [0.262] último [0.286]
primitivo	trueque [0.149]
principal	problema [0.263]
principio	aplicación [0.297] trato [0.299] establecer [0.29]
privación	seno [0.115]
privado	público [0.19]
pro	garantizar [0.017] aumento [0.002] valor [0.001] partido [0.001]
probabilidad	ocurrencia [0.108]
problema	principal [0.263] solución [0.21] político [0.277]
producción	producto [0.344] aumento [0.299] trabajo [0.309] costo [0.288] suma [0.128] valor [0.314] empresa [0.293] bajo [0.263] capital [0.318] consumo [0.29] precio [0.304]
producción-inmaterial	analogía [0.102]
producción-nacional	bruto [0.201]
producto	trabajo [0.318] precio [0.306] producción [0.344] trabajador [0.005] país [0.298] capital [0.317] mercado [0.309] actividad [0.297] valor [0.315] aumento [0.294] todo [0.311] vale [0.058] empresa [0.29] bajo [0.254]
producto-bruto	producto-liquido [0.148]
producto-liquido	producto-bruto [0.148]
producto-marginal	decreciente [0.181]
productor	beneficio [0.251] consumidor [0.29] benefactor [0.272]

Palabra	Vecinos cercanos
profesional	formación [0.213]
profesor	universidad [0.183]
programa	investigador [0.242]
progreso	cien [0.056]
proletario	burgués [0.15]
promedio	anual [0.197] peruano [0.171]
propaganda	acentuar [0.238]
propietario	renta [0.208] tierra [0.196]
propio	actividad [0.296]
protección	frente [0.217] laboral [0.234]
provecho	sacar [0.215]
proveedor	cliente [0.1]
providencia	obligatorio [0.115]
provincia	distrito [0.223] lima [0.191] región [0.146] compensación [0.165] aduana [0.17]
provincial	municipal [0.177]
proyecto	congreso [0.151]
prudencia	precepto [0.144]
prueba	experiencia [0.196]
público	administración [0.165] privado [0.19] sector [0.222] ministro [0.024] estado [0.28] político [0.274] gasto [0.253] recursos [0.283]
puente	carretera [0.154]
puerto	cierre [0.071] carretera [0.088]
puesto	pago [0.153]
química	flor [0.083]
quinto	cuarto [0.087] población-mundial [0.117]
racional	orden [0.221] instinto [0.171] comportamiento [0.25] decisión [0.24]
razón	valer [0.116]
realista	instrumental [0.147]
realizar	actividad [0.286] persona [0.289]
recaudación	impuesto [0.131] descentralización [0.17] monto [0.192] evasión [0.14] tributario [0.165]

Palabra	Vecinos cercanos
reciente	industrialización [0.124]
reciprocidad	mutuo [0.167]
recomendar	implementación [0.136]
recursos	público [0.283] necesidad [0.283] persona [0.283]
recursos-natural	explotación [0.148] ambiente [0.159] agotar [0.132]
red	infraestructura [0.182] usuario [0.133]
reducción	número [0.257]
reembolso	adelante [0.126]
referente	historia [0.19] escenario [0.137]
reforma	constitucional [0.128] estructural [0.156]
refuerzo	castigo [0.162] premio [0.214]
región	regional [0.215] provincia [0.146]
regional	local [0.202] región [0.215]
registro	clave-pública [0.094] certificado [0.163]
regla	excepción [0.136]
reglamento	fondo-estructural [0.082]
regular	normal [0.221]
reino	unido [0.072]
relativo	absoluto [0.245]
renta	impuesto [0.279] tierra [0.252] propietario [0.208] total [0.243]
reparar	daño [0.139]
repartición	dotación [0.147]
reposición	extracción [0.212] etapa [0.214]
represa	rió [0.154] sano [0.063]
representa	cifra [0.201]
república	federal [0.142] presidente [0.118]
res	carne [0.046] familia [0.022] ingreso [0.007]
reservas	banco-central [0.203] divisas [0.181]
respaldar	riqueza-total [0.166]
respaldar-indev	riqueza-total [0.149]
respuesta	pregunta [0.216]
restricción	imponer [0.197]

Palabra	Vecinos cercanos
restricción	tentación [0.089]
resultado	vale [0.076] político [0.298] mero [0.021]
resumir	siguiente [0.126]
reunión	experto [0.17]
rico	pobre [0.246]
rió	represa [0.154]
riqueza	aumento [0.274] hombre [0.292] social [0.247] consumo [0.291] genero [0.284] valor [0.246]
riqueza-artificial	riqueza-natural [0.161]
riqueza-natural	riqueza-artificial [0.161] zona [0.196]
riqueza-propia	isla [0.188]
riqueza-total	zona [0.16] respaldar-indev [0.149] respaldo [0.166]
ritmo	crecimiento [0.093] pib-per-capita [0.088]
rival	manufactura [0.125]
robo	actual [0.011]
rol	economía [0.014] social [0.034]
sacar	provecho [0.215]
saldo	sector-privado [0.126]
salida	entrada [0.158]
salud	educación [0.223] vivienda [0.218] acceso [0.215]
salvaje	deflación [0.118]
sanidad	alimentario [0.125]
sano	representar [0.063]
secreto	criptografía [0.115]
sector	público [0.222]
sector-privado	saldo [0.126]
secundario	primario [0.139]
sede	inseguro [0.054]
segundo	primero [0.262] lugar [0.25] primero [0.269]
seis	cinco [0.161]
semana	junio [0.091] hora [0.108]
semilla	agricultor [0.08]

Palabra	**Vecinos cercanos**
seno	privación [0.115]
sentido	comunismo [0.285]
servicio	préstamo [0.225]
siembra	cosecha [0.131]
siglo	década [0.211]
siguiente	resumir [0.126] cuadro [0.113]
símbolo	movilizar [0.064]
sindical	convenio [0.199]
sindicato	confederación [0.195]
sistema	utilizar [0.288] político [0.29]
sistema-financiero	argentina [0.162]
situación	tipo [0.29]
situar	tipo [0.014]
soberano	territorial [0.146] alimentario [0.124]
social	individual [0.173] trabajo [0.303] vida [0.17] ocio [0.005] político [0.319] economía [0.236] rol [0.034] partido [0.317] sociedad [0.25] desarrollo [0.318] ciencia [0.189] estado [0.303] valor [0.278] riqueza [0.247] derecho [0.226]
sociedad	vida [0.252] social [0.25] individuo [0.282]
socio	economía [0.005] trabajo [0.009] peligro [0.098] condición [0.025] posibilidad [0.041]
solicitar	préstamo [0.168]
solidaridad	agencia [0.096] economía-social [0.168]
solución	problema [0.21] equilibrio [0.236]
soviética	unión [0.209]
stock	new-york [0.052]
subasta	centralizado [0.184] modelo-de-competencia-perfecta [0.142]
subsidiario	filial [0.11]
subsidio	cruz [0.082]
suma	producción [0.128]
superávit	balanza-comercial [0.144] primario [0.194]
superficie	cosecha [0.077]

Palabra	Vecinos cercanos
superior	educación [0.225] inferior [0.155]
supremo	decreto [0.165]
sur	África [0.116] norte [0.196] pacifico [0.063]
sustitución	paradigma [0.173]
sustituto	complementario [0.089]
talento	habilidad [0.18]
tarea	político [0.018]
tarjeta	dinero-electrónico [0.165]
tasa	cambio [0.156] inflación [0.22]
tasa-de-crecimiento	convergente [0.121]
tasa-de-interés	Japón [0.104]
tasa-de-producción	función-de-demanda [0.147]
tasa-marginal-de-sustitución	canasta [0.199] manzana [0.217]
tecnología	nuevo [0.214] desarrollo [0.203] conocimiento [0.252] información [0.238] logística [0.096]
telecomunicaciones	informática [0.178] ingeniería [0.141]
teléfono	televisión [0.07] móvil [0.053]
televisión	teléfono [0.07]
tema	político [0.126] foro [0.029]
tendencia	creciente [0.217]
tensión	línea [0.094]
tentación	restricción [0.089]
teorema	mano [0.152]
teoría	vale [0.082] economía [0.163] general [0.284] modelo [0.258] consideración [0.285] hecho [0.283] explicación [0.254] hipótesis [0.193] científico [0.226] objeto [0.272]
teórico	juego [0.221] empírico [0.219] hipótesis [0.26]
tercer	mundo [0.11]
tercer	bloqueo [0.059]
término	general [0.252] expresar [0.171]
terreno	cultivo [0.185]

Palabra	Vecinos cercanos
territorial	soberano [0.146]
territorio	multinacional [0.179]
tiempo	energía [0.143] información [0.272]
tierra	renta [0.252] propietario [0.196] cultivo [0.143] capital [0.162]
timidez	new-york [0.188] paúl [0.146]
tipo	situar [0.014] situación [0.29]
tipo-de-cambio	nominal [0.174] dólar [0.202] variación [0.187] devaluación [0.235] fijo [0.207] banco-central [0.207]
todo	producto [0.311] foro [0.319] país [0.288] estado [0.314] caso [0.31] objeto [0.301]
toma	decisión [0.216]
torno	gira [0.086] controversia [0.077] postura [0.115]
total	renta [0.243] importación [0.252] millones-de-dólares [0.131]
trabajador	industria [0.017] capital [0.004] país [0.003] individual [0.016] circular [0.039] producto [0.005] necesario [0.008] era [0.016]
trabajo	producto [0.318] aumento [0.258] capital [0.316] producción [0.309] economía [0.262] hombre [0.242] social [0.303] actividad [0.275] país [0.296] empleo [0.216] empresa [0.268] mente [0.026] ocio [0.004] partido [0.318] bajo [0.22] socio [0.009]
tradición	instinto [0.154]
tradicional	instinto [0.186]
tráfico	droga [0.156]
transferencia	distrito [0.15]
transformación	acto [0.057] benefactor [0.234]
transitorio	consejo [0.07]
trato	principio [0.299] establecer [0.294]
treinta	años [0.058] depresión [0.141] cinco [0.126]
tribunal	arbitraje [0.171]
tributario	recaudación [0.165]
trigo	harina [0.145] pan [0.121]
trimestre	cuarto [0.104]
triste	wall-street [0.064]

Palabra	**Vecinos cercanos**
trueque	primitivo [0.149]
tumba	amazonas [0.182]
turismo	turista [0.1]
turista	atracción [0.098] imagen [0.09] turismo [0.1]
último	primero [0.286] instancia [0.074] foro [0.013]
único	ex [0.01]
unido	reino [0.072]
uniforme	marketing [0.139]
unilateral	arancelario [0.158]
unión	aduanero [0.155] soviética [0.209]
universidad	profesor [0.183] entorno [0.185]
urbano	gestión [0.174]
Uruguay	carne [0.14]
usual	lenguaje [0.127]
usuario	red [0.133] certificado [0.22]
útil	marginal [0.158]
utilizar	dinero [0.281] sistema [0.288] método [0.213]
vacío	lleno [0.044]
vale	oro [0.189] producto [0.058] resultado [0.076] teoría [0.082] capital [0.055] razón [0.116]
valor	producción [0.314] producto [0.315] mercancía [0.214] dinero [0.263] social [0.278] moneda [0.209]precio [0.298] ceso [0.011] pro [0.001] cambio [0.284] capital [0.295] necesidad [0.27] riqueza [0.246]
variación	tipo-cambio [0.187]
variedad	novedad [0.088]
vegetal	animal [0.084]
vehículo	gasolina [0.093] mover [0.067]
veinte	cientos [0.169] diez [0.172] años [0.093]
vendedor	comprador [0.307]
vender	comprar [0.287]
ventaja	comercio [0.236]
verdadero	falso [0.117]

Palabra	Vecinos cercanos
vestido	mueble [0.121]
vez	mayor [0.31] mundo [0.296]
viaje	agencia [0.101]
victima	beneficiario [0.111] anexo [0.059]
vida	social [0.17] sociedad [0.252]
vieja	nueva-economía [0.099]
vino	paño [0.259] Francia [0.125]
visión	clásico [0.21]
vivienda	educación [0.167] alimentación [0.159] construcción [0.15] salud [0.218]
volátil	alta [0.131]
voluntario	ayuda [0.158]
voto	ponderado [0.068]
vuelta	depresión [0.137]
wall-street	bolsa [0.114] triste [0.064]
zona	riqueza-natural [0.196] riqueza-total [0.16] euro [0.088] habitar [0.135]

Tabla F.1: Thesaurus de Economía.