BÚSQUEDA DE COLOCACIONES EN LA WEB PARA SINÓNIMOS DE WORDNET

R. Guzmán-Cabrera[1,3], M. Montes -y-Gómez[2,3], P. Rosso[3]

[1] Facultad de Ingeniería Mecánica, Eléctrica y Electrónica
Universidad de Guanajuato, México.
guzmanc@salamanca.ugto.mx
[2] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
mmontesg@inaoep.mx
[3] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
{rguzman, mmontes, prosso }@dsic.upv.es

Resumen

La Web es sin lugar a dudas el repositorio de información más grande jamás construido por el ser humano. Con más de cuatro mil millones de páginas indexadas por los motores de búsqueda públicos, la Web representa el mayor y más amplio corpus textual disponible en la actualidad. Por su valor lingüístico, dado que contiene información en más de 1500 lenguajes, este corpus está siendo usado con gran éxito en muchas tareas de procesamiento del lenguaje natural. En particular, varios métodos de minería de datos se han aplicado para extraer de la Web algunos tipos de patrones lingüísticos útiles para tareas como la traducción automática y búsqueda de respuestas. En este artículo presentamos un método que permite encontrar combinaciones de palabras significativas a los diferentes sentidos atribuibles a una palabra polisémica. Los experimentos realizados, aunque preliminares, muestran el gran potencial del método propuesto para encontrar estas colocaciones por sentido usando la Web como corpus, así como la viabilidad de la incorporación de dichas colocaciones en sistemas de desambiguación del sentido de las palabras, que pueden a su vez ser usados en sistemas de traducción automática y recuperación de información.

Palabras clave: Procesamiento de lenguaje natural, desambiguación del sentido de las palabras, minería de la Web, colocaciones.

SEARCHING COLLOCATIONS ON THE WEB FOR WORDNET SYNSETS

Abstract

There is not any doubt that the Web is the biggest information repository never constructed by the human being. With more than four billion pages indexed by the public search engines, the Web represents the biggest and the widest textual corpus available in our days. Because its high linguistic value, it contains information in more than 1500 different languages, this corpus is being used in many tasks of Natural Language Processing with great success. In particular, several methods of data mining have been applied to extract from the Web some types of linguistic patterns useful for diverse tasks such as automatic translation and question answering. In this work we present a method that allows finding significant word combinations to the different senses attributable to a polysemic word. The experiment results, even preliminary, show a great potential of the proposed method to find these sense collocations by using the Web as a linguistic corpus, as well as the feasibility of incorporation the lexical patterns obtained in word sense disambiguation systems that can be used, for example, in machines translation or information recovery systems.

Keywords: Natural language processing, word sense disambiguation, Web mining, collocations.

1. Introducción

Sin lugar a dudas el tesoro más valioso de humanidad es el conocimiento. Gran parte de este conocimiento existe en forma de lenguaje natural: libros, periódicos, artículos y por supuesto en páginas Web. La posesión real de todo este conocimiento depende de nuestra capacidad para hacer ciertas operaciones con la información textual, por ejemplo: buscar información interesante, comparar fuentes de información diferentes, resumir grandes conjuntos de documentos y traducir textos a diferentes lenguajes.

El Procesamiento de Lenguaje Natural (PLN) es un área de investigación de las ciencias computacionales que se enfoca principalmente en el diseño de mecanismos que permitan a las computadoras comprender el lenguaje natural, así como en el desarrollo de métodos relacionados con las tareas de acceso y análisis de información textual.

Por otra parte, con la denominada sociedad de la información, día a día se multiplica la cantidad de información disponible, lo que conlleva un aumento en la dificultad de procesar esta información con los métodos clásicos. Así, en fechas recientes han surgido técnicas computacionales especializadas en el análisis de grandes cantidades de datos. Una de estas técnicas es la minería de datos (MD), cuya idea clave es que los datos contienen más información de la que se ve a simple vista. Su propósito es extraer información implícita, previamente desconocida y potencialmente útil a partir de los datos (Frawley, 1992). La combinación de técnicas de PLN y MD para el análisis de grandes conjuntos de información textual se conoce como Minería de Texto (MT). La MT aplicada al análisis de conjuntos de páginas Web se denomina minería de contenido de la Web (Etzioni, 1996).

El presente trabajo se enfoca en la minería de contenido de la Web. En él se propone un método que, usando la Web como corpus[1], permite extraer un conjunto de colocaciones atribuibles a cada sentido de una palabra dada, es decir, un conjunto de combinaciones de palabras relacionadas con cada uno de los sentidos de una palabra polisémica. Por ejemplo, para la palabra *banco*, el método propuesto permite descubrir palabras como *dinero*, *depósito* y *tarjeta de crédito* vinculadas al

[1] Un corpus es una colección de texto, en formato electrónico, que puede ser procesado con un equipo de cómputo para varios propósitos, como son, la investigación lingüística y la ingeniería de lenguaje (Kilgarriff, 2003).

sentido de institución financiera, y palabras como *cocina*, *subir* y *mesa* relacionadas con el sentido de mueble.

El uso de la Web como corpus no es una idea novedosa. Desde hace algún tiempo, la Web se ha empleado en el estudio de las lenguas a partir de sus ejemplos de uso (i.e. lingüística de corpus). En particular, en el área de PLN se han aplicado exitosamente técnicas de MT para extraer de la Web distintos tipos de patrones lingüísticos especialmente útiles para la traducción automática (Grefenstette, 1999), el aprendizaje de ontologías (Aguirre, 2000), la búsqueda de respuestas (Brill, 2001), la resolución de anáfora (Bunescu, 2003) y la desambiguación del sentido de las palabras (Aguirre, 1995; Montoyo, 2000; Rosso, 2003).

El uso de la Web como corpus ha sido motivado principalmente por su tamaño y diversidad de lenguajes. En julio de 1999 el tamaño de la Web se estimaba en 56 millones de direcciones, 125 millones en enero del 2001 y 172 millones en enero del 2003. Se puede apreciar un enorme crecimiento de más del 200% en poco menos de 5 años. Por otra parte, en 1999 se encontraron 800 millones de páginas Web indexadas; si estimamos que el tamaño promedio de una página Web es de 8 Kilobytes de texto sin formato, tendremos entonces cerca de 6 Terabytes (1 TB = 1024 Gigabytes) de texto disponible en 1999 y aproximadamente 30 Terabytes en el 2003. Estas cifras indican claramente que la Web es un corpus inmenso. Además, la Web es un repositorio de información multilingüe, ya que aproximadamente 71% de las páginas están escritas en inglés, 6.8% en japonés, 5.1% en alemán, 1.8% en francés, 1.5% en chino, 1.1% en español, 0.9% en italiano, 0.7% en sueco, y el restante 11.1% está repartido en otros idiomas y dialectos (Kilgarriff, 2003).

Sin embargo, y a pesar de todas estas bondades, la Web tiene varios aspectos negativos. Entre estos destacan los siguientes: su información es muy heterogénea y desorganizada, existe mucha información irrelevante, la información esta etiquetada de diferentes formas dificultando su procesamiento, y finalmente y no menos importante, la información no es validada por ningún organismo y por tanto puede existir mucha información incorrecta.

La tarea de encontrar automáticamente relaciones semánticas entre palabras contiguas ha atraído la atención de muchos investigadores del área de PLN en las últimas décadas (Celina, 2003; Mihalcea, 2004; Smandja, 1993). Como resultado de estas investigaciones se han escrito

importantes diccionarios de colocaciones[2] en inglés, por ejemplo, el creado por Benson, (Benson, 1989). Además, se han desarrollado sistemas que permiten hacer análisis de colocaciones. Un ejemplo de estos sistemas es el N-Gram Statistics Package[3] desarrollado por Ted Petersen, que permite analizar estadísticamente las palabras de un corpus.

El trabajo que aquí se presenta, al igual que otros previamente mencionados, se enfoca en la extracción de patrones lingüísticos (combinaciones de palabras de uso común) a partir del uso de la Web como corpus y la aplicación de técnicas estadísticas para su análisis. Además, aprovechamos la redundancia de la información en la Web para discernir entre la información correcta y la incorrecta. Sin embargo, a diferencia de estos trabajos, el nuestro se distingue por:

- Extraer colocaciones significativas para los diferentes sentidos atribuibles a una palabra dada, en lugar de simplemente extraer colocaciones significativas a un dominio o lenguaje especificado.
- Usar la Web para extraer conocimiento (i.e., las colocaciones previamente citadas) directamente aplicable en la desambiguación del sentido de las palabras, en lugar de solamente seleccionar ejemplos de uso para los distintos sentidos de una palabra.
- Aplicar nuevas medidas para la evaluación de la relevancia de las colocaciones identificadas, las cuales no sólo permiten establecer su carácter recurrente y poco aleatorio, sino también su vinculación a un sentido particular de la palabra en cuestión.

Las colocaciones extraídas mediante el método propuesto tendrán una aplicación directa en la tarea de Desambiguación del Sentido de las Palabras (WSD, Word Sense Disambiguation en inglés). Esta tarea consiste en asociar una palabra, inmersa en un contexto de uso dado, con uno de sus posibles significados (Aguirre, 1995). Aunque la tarea de WSD no es un fin en sí misma, es una etapa indispensable para el análisis sintáctico y la interpretación semántica en tareas de PLN. Además es un módulo importante en aplicaciones de recuperación de información (Kurt, 2000), clasificación de textos (Kosala, 2000) y traducción automática (Smrz, 2001) entre otras. Por ejemplo, un sistema de recuperación de información tradicional, al cual se somete la pregunta ¿que plantas viven en el desierto?, responderá con todos los documentos que contienen los términos

[2] Una colocación es una combinación arbitraria y recurrente de palabras.
[3] http://www.d.umn.edu/~tpederse/code.html

plantas y desierto independientemente de su significado. En algunos documentos el término planta tendrá el sentido de “ser vivo”, y en otros el de “industria”. Mientras que un sistema de recuperación de información que emplee un módulo de WSD será capaz de distinguir entre los sentidos de los términos de la consulta y sólo devolverá los documentos en los que se usa la palabra planta con el sentido de ser vivo.

Lo que resta del artículo se organiza de la siguiente manera. La sección 2 describe el método utilizado para encontrar en la Web las colocaciones significativas para una palabra dada. La sección 3 presenta los resultados preliminares de los experimentos realizados y algunos ejemplos de los patrones léxicos encontrados para cada uno de los sentidos de la palabra instance. Finalmente, la sección 4 expone nuestras conclusiones y el trabajo futuro.

2. Metodología

En el lenguaje natural hay muchas combinaciones de palabras que ocurren con frecuencia y corresponden a un uso particular de una palabra o un sentido de una frase. Actualmente, para una palabra polisémica (con varios posibles significados) se distinguen dos tipos de combinaciones de palabras relevantes: asociaciones y secuencias léxicas.

Las asociaciones léxicas son simplemente un conjunto de palabras que ocurren frecuentemente y de manera simultánea, pero sin un orden preestablecido, con un sentido particular de la palabra en cuestión. Por su parte, las secuencias léxicas son cadenas de palabras, que tienen un orden fijo, y que también se relacionan con un sentido único de la palabra en cuestión.

El método que se describe a continuación permite reconocer ambos tipos de colocaciones para una palabra polisémica dada.

1. **Determinar los sentidos atribuibles a la palabra dada:** Para ello debe consultarse una base de datos léxica con dicha información. Para el caso del inglés la base de datos léxica comúnmente usada es *WordNet* (Fellbaum, 1998), la cual es una base de datos léxico-conceptual estructurada en forma de red semántica e inspirada en teorías psicolingüisticas sobre la memoria léxica humana. WordNet almacena información sobre palabras pertenecientes a las categorías sintácticas de sustantivo, verbo, adjetivo y adverbio.

2. **Obtener el conjunto de sinónimos relativos a cada sentido:** Para cada sentido de la palabra, indicado en la base de datos léxica utilizada, se obtiene su conjunto de sinónimos. En el caso de WordNet las palabras se organizan en conjuntos de sinónimos (*synsets*), cada uno de los cuales representa un concepto léxico diferente. Cada *synset* contiene la lista de palabras sinónimas, además de información de relaciones semánticas establecidas con otras palabras o *synsets*.
3. **Descargar snippets relacionados con cada uno de los sentidos:** Usando como patrón de búsqueda en la Web cada uno de los sinónimos se usa un motor de búsqueda (por ejemplo Google) para descargar un conjunto de snippets. Los snnipets son los pequeños resúmenes que los motores de búsqueda suelen incluir como información de cada una de las páginas retornadas. En nuestro caso los snippets representan expresiones de uso común de los sinónimos de la palabra polisémica dada.
4. **Construir un corpus por sentido:** Los snippets de todos los sinónimos de un sentido son filtrados y agrupados, conformando con ello un corpus por sentido de la palabra polisémica dada. Este corpus por sentido es construido considerando las diez palabras alrededor del sinónimo, cinco a la derecha y cinco a la izquierda. Esta decisión respecto al tamaño del contexto se basó en las experiencias previas en WSD, principalmente en los resultados obtenidos en la competencia SENSEVAL[4].
5. **Extraer las colocaciones relevantes a cada sentido:** En primer lugar se encuentran *todas* las colocaciones (tanto asociaciones como secuencias) posiblemente relacionadas con cada uno de los sentidos de la palabra dada. Para ello se aplican algoritmos tradicionales para el descubrimiento de asociaciones y secuencias (Agrawal, 1994) sobre los corpora por sentido. En segundo lugar se aplican los siguientes criterios para la selección de las colocaciones más representativas para cada uno de los sentidos:
    a. ***Fuerza*.-** una colocación, secuencia o asociación, es relevante si es frecuente, esto es si ocurre un número de veces mayor a un valor de umbral o frecuencia de corte determinado previamente; y está definida por:

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

$$\frac{f - \bar{f}}{\sigma} \geq umbral \qquad \text{donde} \quad \begin{cases} f & \text{es la frecuencia de la palabra en cuestión} \\ \bar{f} & \text{es la frecuencia promedio y} \\ \sigma & \text{es la desviación estandar} \end{cases}$$

El umbral está definido en un valor igual a la suma de la frecuencia promedio y la desviación estándar, y es la frecuencia mínima que deben tener las palabras para superar esta medida. Con lo cual aseguramos la extracción sólo de aquellas ocurrencias que aparecen de manera recurrente en los contextos del sentido de la palabra, eliminando todas las palabras que pudieran aparecer de manera casual.

b. ***Dispersión interna***.- Es una medida binaria que indica si la colocación –asociación o secuencia– se presentó con todos los sinónimos pertenecientes al sentido en cuestión. Esta medida asegura que las colocaciones seleccionadas sean representativas a nivel sentido de WordNet, pues las colocaciones que no están en el contexto de todos los sinónimos que componen al sentido de la palabra, según WordNet, son descartadas.

c. ***Dispersión externa***.- Es una medida binaria que indica si la colocación es exclusiva de un sentido de la palabra o ocurre con varios sentidos. La idea es seleccionar solamente aquellas asociaciones y secuencias vinculadas a un único sentido de la palabra según WordNet.

La metodología recién descrita puede ser aplicada para extraer tanto asociaciones como secuencias relacionadas con los sentidos de las palabras. A continuación se explica brevemente algunos detalles técnicos para la extracción de ambos tipos de colocaciones.

Extracción de asociaciones.- En primer lugar, con ayuda de la base de datos léxica WordNet, se identifican los sentidos admisibles para la palabra dada, así como el conjunto de sinónimos (synset) relacionado con cada sentido. Después, en la Web se identifican algunos ejemplos de uso de cada sinónimo, y las palabras del contexto se introducen en una lista, llevando un conteo de las ocurrencias de cada una de ellas. A partir de la lista se seleccionan aquellas palabras de contexto que superan las medidas de fuerza y dispersión mencionadas en el punto 5 de la metodología. Cabe mencionar que las palabras que se encuentran en la tabla resultante no es condición que se encuentren de manera contigua al sinónimo. Su posición dentro del contexto es variable dentro de la ventana definida, de esta manera encontramos aquellas palabras que se encuentran vinculadas de manera significativa con el sentido. Finalmente, si se desea encontrar asociaciones conformadas por

más de una palabra de contexto se aplica sobre la lista de palabras obtenida el proceso descrito en (Bayardo, 1999) para identificación de los conjuntos frecuentes de palabras.

Extracción de secuencias.- La extracción de las secuencias, al igual que la de asociaciones se basa en el uso de WordNet como base de datos léxica y en la conformación de un corpus por sentido a partir de snippets bajados de la Web. La extracción de estas secuencias de palabras se hace automáticamente siguiendo un proceso iterativo que considera diferentes tamaños de ventana alrededor de la palabra dada, desde una ventana de tamaño 1 hasta una de tamaño 5. Para cada tamaño de ventana se toman las palabras respetando su ubicación respecto al sinónimo de la palabra estudiada y se insertan en una lista considerando su frecuencia en el corpus de referencia. Después, al igual que para las asociaciones léxicas, las secuencias resultantes son filtradas y sólo aquellas que son significativas al sentido de WordNet se mantienen. En este caso, a diferencia de las asociaciones, las palabras que forman la secuencia son contiguas entre ellas, y respetan su posición dentro del contexto.

3. Resultados preliminares

En esta sección se muestran los resultados obtenidos al aplicar la metodología descrita en la sección 2 para el análisis de la palabra *instance*. Elegimos esta palabra, ya que además de ser polisémica tiene dos sinónimos comunes entre sus sentidos, lo que nos permitirá exhibir el efecto de las medidas de dispersión en los patrones léxicos obtenidos. Los sentidos marcados en WordNet para la palabra *instance*, en su función de sustantivo, son los siguientes:

1. *case, instance, example -- an occurrence of something*, (una ocurrencia de algo).
2. *example, illustration, instance, representative -- an item of information that is representative of a type*, (un artículo de información que es representativo de un tipo).

Usando como patrón de búsqueda en la Web los sinónimos de cada sentido se bajaron snippets (para ello se uso el motor de búsqueda Google), los cuales contienen ejemplos de contexto de uso de cada sentido, esto es, expresiones de uso común de los sinónimos que componen al sentido de WordNet correspondiente. El número de snippets bajados por sinónimo se muestra en la tabla 1.

Con estos snippets se formaron 2 corpus, uno para cada sentido de *instance*. El corpus para el primer sentido fue formado por la unión de 2,826 snippets, mientras que el del segundo sentido se formó con 3,881 snippets. Para el segundo sentido se tienen más snippets dado que contiene un sinónimo más. En la tabla 2 se muestra el resumen de los resultados obtenidos en la extracción de asociaciones. En esta tabla es evidente el nivel de restricción impuesto por los criterios de selección de las colocaciones relevante. Por ejemplo, para el primer sentido de *instance*, se encontraron 12,684 ejemplos de uso. A partir de éstos se identificaron 2,831 palabras distintas en el contexto inmediato de los sinónimos. De estas palabras solamente 179 superaron el umbral de frecuencia definido (frecuencia mayor a la frecuencia promedio más la desviación estándar), y sólo 25 cumplieron también con las restricciones de dispersión interna y externa. Estas palabras, asociaciones léxicas simples, se muestran en la tabla 3.

Con el propósito de ejemplificar el uso de las palabras significativas asociadas con los dos sentidos de *instance* se presentan algunas oraciones de uso común. Con ellas se puede intuir el uso de las colocaciones, llámense asociaciones o secuencias, en la desambiguación del sentido de las palabras.

Sentido 1:

…another **instance** of the same **process** already running on the current machine …

... Enforcing a rule that only one **instance** of **process** is running is an interesting task …

Sentido 2:

... Activity in this **instance** involves the use of **government** facilities and equipment for …

… Another difference between **instance members** and class **members** is that class …

En el caso de las secuencias vinculadas a los dos sentidos de la palabra *instance* los resultados obtenidos para diferentes valores de ventana *V* se muestran en la tabla 4. En esta tabla se puede observar que a medida que se incrementa el tamaño de ventana el número de secuencias disminuye, este comportamiento se debe a que la secuencia (de una, dos o más palabras) debe formar parte del contexto de todos los sinónimos de la palabra en cuestión, y además debe mantener el mismo orden de aparición. Para una secuencia que ha superado las medidas de fuerza y dispersión, a mayor frecuencia y mayor tamaño más significativa será.

En la tabla 5 se muestran algunas secuencias relacionadas con los dos sentidos de la palabra *instance*. Las secuencias están separadas por sentido y por su localización izquierda o derecha con respecto al sinónimo, estas secuencias se encuentran en expresiones de uso común como "*customers instance*", "*graphic design instance*", "*instance design*" o "*instance studies case*".

En este experimento no prescindimos de las palabras vacías, como preposiciones y determinantes, pues consideramos que éstas juegan un papel importante en la asignación de un sentido a una frase. Por ejemplo, la preposición *for* la encontramos asociada de manera significativa con el segundo sentido de *instante*, dado que es usada en la expresión de uso común "*for instance*".

4. Conclusiones y trabajo futuro

En el artículo se presentó un método que permite extraer colocaciones significativas relacionadas con cada sentido de una palabra dada a partir de la Web. El método considera la extracción de dos tipos de colocaciones, asociaciones y secuencias. Los resultados presentados, aunque preliminares y de carácter ilustrativo, muestran por una parte el gran potencial de la Web como corpus lingüístico, y por otra parte, la viabilidad de la incorporación de las colocaciones descubiertas en los sistemas de desambiguación del sentido de las palabras.

La principal aportación del trabajo, desde nuestra perspectiva, radica en la determinación de colocaciones significativas para cada sentido de una palabra polisémica. Esto es notablemente diferente respecto a otros trabajos previos, donde las colocaciones son relativas a un dominio o lenguaje especificado. Además se han propuesto un conjunto de criterios especiales para el filtrado de las colocaciones por sentidos. Estos criterios se basan en los conceptos básicos de las técnicas de selección de atributos (en los problemas de clasificación), y nos permitieron no sólo seleccionar las colocaciones mas frecuentes, sino también las mas fuertemente vinculadas a un solo sentido, es decir, las menos ambiguas.

Aunque el ejemplo mostrado en el artículo consideró el análisis de un sustantivo en inglés, nuestra metodología es suficientemente general para aplicarse a palabras de otras categorías

morfosintácticas, así como a cualquier otro lenguaje para el cual se disponga de una base de datos léxica similar a WordNet.

Actualmente las medidas de dispersión interna y externa son muy rígidas, pues filtran las asociaciones y secuencias que no ocurren con todos los sinónimos de un sentido, y en uno solo de los sentidos, independientemente de sus frecuencias. Este esquema de medición causa que muchas colocaciones realmente significativas queden fuera del análisis. Parte del trabajo futuro se encamina a la solución de este problema, incluyendo la frecuencia en la medición de dichas dispersiones.

Asimismo se plantea como parte del trabajo futuro probar las colocaciones extraídas con el método propuesto en la tarea de desambiguación del sentido de las palabras (WSD). Debido a que por el momento no se han obtenido demasiados patrones por sentido consideramos que la mejor manera de integrar estos patrones en la tarea de WSD es a manera de atributos complementarios en un esquema supervisado.

Agradecimientos

Los autores agradecen a las siguientes instituciones y organismos por su ayuda parcial a la realización de este proyecto: PROMEP (UGTO-121), R2D2 CICYT (TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), CONACYT (43990), así como también a la Secretaría de Estado de Educación y Universidades de España.

Referencias

a) Agrawal, R., Srikant, R., 1994, Fast Algorithms for Mining Association Rules, VLDB-94.

b) Aguirre, E., Rigau, G., (1995) A Proposal for Word Sense Disambiguation using Conceptual Distance, *Recent Advances in NLP*, RANLP'95.

c) Aguirre, E., Olatz, A., Hovy M., (2000) Enriching Very Large Ontologies using the WWW, ECAI 2000, *Workshop on Ontology Learning*. Berlin.

d) Bayardo, R., Agrawal, R., (1999), Mining the Most Interesting Rules, Knowledge Discovery and Data Mining, 5th ACM SIGKDD.

e) Benson, M. (1989), *The BBI Combinatory Dictionary of English*. Amsterdan, Philadelphia: John Benjamin Pr.

f) Brill, E., Lin, J., Banko, M., Dumais, S., (2001) Data-intensive Question Answering., Proc. Of the Tenth Text Retrieval Conference TREC-2001.

g) Bunescu, R., (2003), Associative Anaphora Resolution: A Web-Based Approach, EACL-2003 Workshop on the Computational Treatment of Anaphora, Budapest, Hungary.

h) Celina, S., Gonzalo J. and Verdejo F., (2003), Automatic Association of Web Directories with Word Senses, *Computational Linguistics*, 29, 485-502.

i) Etzioni O. (1996) The World Wide Web: Quagmire or Gold Mine?, *Communications of the ACM*,.39, 11, 65-68.

j) Fellbaum, Ch. (1998), *WordNet as Electronic Lexical Database*. MIT Press.

k) Frawley, W. and Piatesky-Shapiro, G., (1992), Knowledge Discovery in Databases: An Overview, *AI Magazine*, pp 213-228.

l) Grefenstette, G., (1999) The World Wide Web as a resource for example-based Machine Translation Task, Proc. Of Aslib Conference on Translating and the Computer. London.

m) Kilgarriff, A. and Greffenstette, G., (2003), Introduction to the Special Issue on Web as Corpus, *Computational Linguistics*, 29, (3), 1-15.

n) Kosala, R. y Blockeel, H., (2000) Web Mining Research: a survey, *SIG KDD Explorations*, 2, 1-15.

o) Kurt, D., Bollacker, S., Lee C., (2000) Discovering Relevant Scientific Literature on The Web, IEEE Intelligent Systems, Volume 15, Number 2, pp. 42–47.

p) Mihalcea, R.,(2004) Making Sense Out of the Web, *Workshop on Lexical Resources and the Web for Word Sense Disambiguation, IBERAMIA, Mexico, 2004.*

q) Montoyo, A., (2000), Metodo Basado en Marcas de Especificidad para WSD. Procesamiento de Lenhuaje Natural n. 26.

r) Rosso, P., Masulli, F., Buscaldi, D., Pla, F., Molina, A., (2003) Automatic Noun Disambiguation, Lecture Notes in Computer Science, vol. 2588. Springer-Verlag.

s) Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*. 7(4):143–177.

t) Smrz, P., (2001) Finding Semantically Related Words in Large Corpora, *FIMU Report Series*:Masaryk University.

Tabla 1

	Sinónimo	Snippets
Sentido 1	case	919
	instance	924
	example	983
Sentido 2	illustration	987
	representative	987

Tabla 2

Palabra: Instance	Sentido1	Sentido 2
Número de ejemplos de uso en el corpus	12,684	15,848
Número de palabras distintas	2,831	3,590
Ocurrencia media de las palabras	4.5	4.4
Desviación estándar de la ocurrencia	7.3	7.5
Número de palabras que superan el umbral de frecuencia	179	238
Palabras que superan la dispersión interna (además del umbral de frecuencia)	87	67
Palabras que superan la dispersión externa (además de la frecuencia y dispersión interna)	25	9

Tabla 3

	Asociaciones léxicas
Sentido 1	based, case, code, data, date, definition, documents, example, examples, file, index, instance, It, java, learning, multiple, net, number, org, our, process, proposal, server, use, will.
Sentido 2	english, free, government, library, link, members, resources, section, software.

Tabla 4

	Sentido 1		Sentido 2	
V	Descubiertas	Significativas	Descubiertas	Significativas
-3	20	3	35	5
-2	29	6	60	12
-1	62	14	100	20
1	72	18	140	28
2	57	8	89	18
3	23	3	61	9
total	263	52	485	92

Tabla 5

		Secuencias Significativas
Sentido 1	Izquierda	customers, home, name, party, resources, this, to, tools, bottle, us party , the bottle, studies customers, to the, the us party, in the bottle.
	Derecha	design, edu, index, law, studies, western, study, code, western reserve, studies case, studies catalogs, studies in, western reserve university, studies catalogs resources.
Sentido 2	Izquierda	Design, art, bouchard, medical, for, fanny bouchard, graphic design, page layout, visual arts, multimedia design, proactive core component, of the mouth.
	Derecha	and, clients, the, in , is, of the, of a, in the, employees post a , of judicial activism, , and fine art , of the secretary, government by john, of the mouth, , , , , , , , , ,