INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Recuperación de información con resolución de ambigüedad de sentidos de palabras para el español

T E S I S

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN
PRESENTA

YOEL LEDO MEZQUITA

Director:
Dr. Grigori Sidorov

Codirector:
Dr. Alexandre Guelboukn

INSTITUTO POLITÉCNICO NACIONAL

SIP-14

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D. F. siendo las 18:00 horas del día 20 del mes de Enero de 2006 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis de grado titulada:

"RECUPERACIÓN DE INFORMACIÓN CON RESOLUCIÓN DE AMBIGÜEDAD DE SENTIDOS DE PALABRAS PARA EL ESPAÑOL"

Presentada por el alumno:

LEDO	**MEZQUITA**	**YOEL**
Apellido paterno	materno	nombre(s)

Con registro:

B	0	1	1	4	0	1

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

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

LA COMISIÓN REVISORA

Presidente

DR. AGUSTÍN FRANCISCO GUTIÉRREZ TORNÉS

Secretario

DR. LUIS PASTOR SÁNCHEZ FERNÁNDEZ

Primer vocal

DR. GRIGORI SIDOROV

Segundo vocal

DR. ALEXANDRE GUELBOUKH KAHN

Tercer vocal

DRA. SOFÍA NATALIA GALICIA HARO

EL PRESIDENTE DEL COLEGIO

DR. HUGO CÉSAR COYOTE ESTRADA

INSTITUTO POLITECNICO NACIONAL SECRETARIA DE INVESTIGACION Y POSGRADO 10 FEB 2006 RECIBIDO DIRECCION POSGRADO

INSTITUTO POLITECNICO NACIONAL CENTRO DE INVESTIGACION EN COMPUTACION DIRECCION

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE CESIÓN DE DERECHOS

En la ciudad de México D. F. el día 07 del mes de junio del año 2003, el que suscribe **Yoel Ledo Mezquita**, alumno del Programa de Doctorado en Ciencias de la Computación con número de registro: **B011401** Adscrito al Centro de Investigación en Computación del IPN, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. Grigori Sidorov y como Co-director al Dr. Alexandre Guelboukh y cede los derechos del trabajo intitulado **Recuperación de información con resolución de ambigüedad de sentidos de palabras para el español** al Instituto Politécnico Nacional para su difusión, con fines académicos y de Investigación.

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

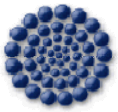
Yoel Ledo Mezquita
Nombre y firma

Apoyos institucionales

Los resultados de esta tesis han sido apoyados en

- **México por:**

Gobierno de México



Consejo Nacional de Ciencia y Tecnologia (CONACyT)

Sistema Nacional de Investigadores (SNI)

Secretaría de Relaciones Exteriores (SRE)

Instituto Politécnico Nacional (IPN)

SIP

Secretaría de Investigación y Posgrado (SIP) del IPN

Comisión de Operación y Fomento de Actividades Académicas (COFAA) del IPN

PIFI

Programa Institucional de Formación de Investigadores (PIFI) del IPN

Centro de Investigación en Computación (CIC) del IPN

Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC

Centro de Investigacion Científica y de Educación Superior de Ensenada (CICESE)

Dirección de Telemática del CICESE

Universidad de las Américas A.C. (UDLA).

Tecnológico de Monterrey (ITESM). Campus Ciudad de México. (CCM).

Universidad del Valle de México. (UVM).

- **Europa por:**

Ministerio de Educación y Ciencia del Gobierno de España

Ministerio de Asuntos Exteriores y de Cooperación del Gobierno de España

Agencia Española de Cooperación Internacional

Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED)

Red Iberoamericana de Tecnologías de Software para la Década del 2000 (RITOS2). Subprograma VII Electrónica e Informática Aplicada del CYTED

- **Cuba por:**

Gobierno de Cuba

Ministerio de Educación Superior (MES)

Instituto Superior Politécnico José Antonio Echeverría (Cujae)

Departamento de Telemática de la Cujae

Agradecimientos

Para la culminación exitosa de este trabajo han dedicado gran parte de su tiempo muchas personas, a los cuales, les agradezco eternamente su cordial ayuda y consejos.

Algunos de ellos normalmente los tengo bien de cerca, tan cerca, que pudiera empezar, por mi Ivette, a la cual le debo gran estimulación a terminar estos estudios.

A mi mamá, Hortensia, y mi papá, Nelson, que desde hace mucho me están ayudando.

Al resto de mi familia, tan unida y que siempre se ha sacrificado para que yo me supere y salga adelante.

A mis Directores de Tesis, los Doctores Grigori Sidorov y Alexandre Guelboukn que sin sus consejos, experiencias, exigencias y paciencia, jamás hubiera terminado este trabajo.

A mis sinodales los Doctores Agustín Gutiérrez, Sofía Galicia y Luis Sánchez por sus consejos y exigencia.

A todos mis amigos y compañeros en Cuba, México, España, Estados Unidos y donde quiera que se encuentren, por su apoyo y confianza.

Dedicatoria

A mis familiares y amigos.
A mis profesores y alumnos.

Abstract

From the beginning of information digital services, information retrieval has drawn researchers´ attention concerning knowledge management, artificial intelligence, computer systems and computer linguistics. It is important because of the need of efficiency in the searching algorithm, in order to find accurate results to requested information rapidly within a great diversity of texts.

One of the problems regarding the portals of information retrieval on Internet (dynamic portals of Altavista, Google, Yahoo!, etc.) and digital libraries (Library of the Congress and others), is the overabundance of different and irrelevant answers. For example:

- A banker searches for “bank” and he obtains answers on “institution for saving and borrowing money”, “river bank”, “heap of snow”, “socket for connecting a memory chip in computers”, “phone bank” and many others.
- A geologist searches for “well” and he finds out information about “excavation, hole in the ground”, “an abundant source”, “an enclosed compartment in a ship or plane for holding something”, “in a good or proper or satisfactory manner or to a high standard” and others.

These inaccuracies are due to different meanings of words, which are known as Word Sense Disambiguation (WSD).

This term is a linguistic mechanism that detects the most suitable sense of a word, according to the context where the word is used, based on its possible definitions.

In this work the development sets out **a new method of Word Sense Disambiguation by means of lexical resources.**

Resumen

Uno de los problemas en los portales de recuperación de información en Internet (los portales dinámicos de Altavista, Google, Yahoo!, etc.) y en bibliotecas digitales (Biblioteca del Congreso de los EE.UU., etc.) es el de brindar diversas respuestas con muy baja pertinencia.

Por ejemplo, un mecánico de autos busca "¿dónde comprar un gato?" y obtiene respuestas sobre los "gatos monteses", "gatos siameses", y otros. Un comerciante de frutas busca "producción de lima" y obtiene respuestas sobre la "ciudad de Lima", "jugo de lima", "lima de uñas", y otros. Estas imprecisiones se deben a los distintos significados que tienen las palabras, lo que se conoce como Desambiguación del Sentido de las Palabras (*Word Sense Disambiguation*, WSD, del inglés.)

Este término es un mecanismo lingüístico para definir el sentido correcto de una palabra, basándose en el contexto donde se emplee y en función de sus posibles sentidos semánticos.

Las aportaciones de esta tesis consisten en el desarrollo de **un nuevo Método de desambiguación de sentidos de palabras usando grandes recursos léxicos.**

Contenido general de la tesis

Índice detallado de la tesis

Lista de figuras

Introducción

En este capítulo se justifica la necesidad de trabajar en la recuperación de la información en portales de Internet.

Se plantean los objetivos de investigación de este trabajo, los cuales se centran en el diseño de un Método de desambiguación de sentidos de palabras usando grandes recursos léxicos.

Se describe la organización de la tesis y el contenido de los capítulos subsiguientes.

Introducción

Motivación

Desde el inicio de los servicios digitales de información, la recuperación de la información ha sido punto de atención para los investigadores en gestión del conocimiento, inteligencia artificial, sistemas computacionales y lingüística computacional. Su importancia radica en que la eficiencia del algoritmo de búsqueda garantice el acceso rápido y pertinente de la información solicitada dentro de un entorno en que se procesan grandes volúmenes de texto. Este algoritmo de búsqueda es uno de los problemas importantes a resolver dentro de la gestión del conocimiento y el procesamiento del lenguaje natural (PLN).

En los últimos años, con el desarrollo de los servicios digitales de información de los medios de almacenamiento masivo y de las redes de telecomunicaciones se ha difundido el uso de las bibliotecas digitales con búsquedas en línea, dejando atrás los algoritmos de búsqueda tradicionales e imponiéndose el desarrollo de algoritmos de búsqueda inteligentes que garanticen mejores resultados en las búsquedas temáticas.

Sin embargo, el desarrollo de métodos para a la solución de dicha problemática, no ha sido muy investigado para la recuperación de la información y en la lingüística computacional. Producir conocimiento sobre las búsquedas inteligentes para los servicios en línea, las condiciones que la determinan y los mecanismos de procesamiento basados en el PLN, son las metas inmediatas a lograr.

Cualquier sistema de procesamiento del lenguaje natural necesita utilizar abundante conocimiento sobre las estructuras del lenguaje, las cuales son de tipo morfológico, sintáctico, semántico y pragmático. El conocimiento morfológico

proporciona información de cómo se construyen las palabras; el sintáctico de cómo combinar las palabras para formar oraciones; el semántico sobre lo que significan las palabras y cómo éste contribuye en el significado completo de la oración, y por último, el pragmático sobre cómo el contexto afecta a la interpretación de las oraciones.

Todas las formas anteriores de conocimiento lingüístico tienen un problema asociado: la ambigüedad. Por lo tanto, la resolución de este tipo de problema es uno de los objetivos principales de cualquier sistema de PLN. Se distinguen diversos tipos de ambigüedades: estructural, léxica, de ámbito, de cuantificación, de función contextual y referencial.

El presente trabajo se centra en la resolución de la ambigüedad léxica, la cual aparece cuando las palabras presentan una misma grafía con diferentes significados. A esta tarea se le conoce como desambiguación del sentido de las palabras (*Word Sense Disambiguation*, WSD, del inglés.)

La resolución de la ambigüedad de los sentidos de las palabras es un mecanismo lingüístico para definir el sentido más adecuado de una palabra, según el contexto donde se emplee, que se define en función de los posibles sentidos de las palabras.

Por ejemplo, un mecánico de autos busca *"¿dónde comprar un gato?"* y obtiene respuestas sobre los *"gatos siameses"*, *"gatos monteses"* y otros. Un comerciante de frutas busca *"producción de lima"* y obtiene respuestas sobre la *"ciudad de Lima en Perú", "fruta lima", "herramientas para limar metales.* Estas imprecisiones se deben a los distintos sentidos que tienen las palabras.

Descripción del problema

La recuperación de información (*Information Retrieval*, IR, del inglés) consiste en la tarea de ordenar los documentos, tanto de texto como de multimedia, que pertenecen a una colección dada de acuerdo a la probabilidad estimada de relevancia para las necesidades de información del usuario. Estas necesidades de información se expresan

generalmente por el usuario en función de las respuestas obtenidas a un requerimiento de un lenguaje no formalizado (por ejemplo, sentencias) o un conjunto de términos en un lenguaje natural.

El esfuerzo requerido para la recuperación de la información es notoriamente complejo debido a que la relación de "relevancia" entre los documentos y las necesidades de información son dependientes de las preferencias e interpretaciones subjetivas del usuario. Además, esta relación es no formalizable [Saracevic, 1995].

La enorme disponibilidad actual de documentos almacenados electrónicamente, especialmente en plataformas distribuidas, ha transformado la recuperación de la información en una disciplina importante. La World Wide Web (WWW) contiene grandes cantidades de información *(unos 2000 millones de páginas que abarcan unos 38 terabytes de datos y que crece 7 millones de páginas diariamente-; también contiene alrededor de 450 millones de imágenes, julio de 2000)* [Pimienta, 2000], [Lawrence, 2000] potencialmente interesantes y accesibles para muchos usuarios (*615 millones para el 2002, de los que 48 millones hablan español; 1030 millones para el 2005, de los que 80 millones hablan español)* [Global Reach, 2002].

Estas cifras actualizadas al 2006 por la *University of California at Berkeley* y *Whois.Net Domain-Based Research* Services plantean la existencia de 1,000 millones de usuarios en Internet, que alcanzará los 2 mil millones para el 2016, 530 mil terabytes de información en Internet, 320 millones de búsquedas diarías y más de 95 millones de dominios resgistrados (38 millones en uso continuo).

Uno de los problemas de recuperación de información en los portales de Internet como los portales dinámicos Altavista, Google, Yahoo, etc., y en bibliotecas digitales como la Biblioteca del Congreso de los USA, es el de brindar diversas respuestas con muy baja pertinencia con respecto a los intereses del usuario.

Por ejemplo: un economista busca *"historia del banco"* y obtiene respuestas sobre los *"bancos de arena"*, *"bancos de madera"* y las *"instituciones financieras"*. Un músico busca *"formato de letra"* y obtiene respuestas sobre el *"documento comercial de pago"*, *"letras del alfabeto"* y *"letras musicales"*. Estas imprecisiones se deben a los distintos sentidos que tienen las palabras.

La WSD es considerada como uno de los problemas de investigación más importantes en el procesamiento del lenguaje natural [Wilks y Stevenson, 1996]. Es esencial para las aplicaciones que requieren la comprensión del lenguaje y de mensajes, la comunicación hombre-máquina, la recuperación de información y otros. Se requiere en aplicaciones de:

- **Traducción automática:** se refiere más que nada a la traducción correcta de información de un lenguaje a otro según lo que se quiera expresar en cada oración y no sólo palabra por palabra. Una aproximación a este tipo de traductores en Internet es el Babylon.

- **Extracción de información y resúmenes.** Los nuevos programas deben tener la capacidad de crear el resumen de un documento sobre la base de los datos proporcionados, con un análisis detallado del contenido sin truncar las primeras líneas de los párrafos.
- **Reconocimiento de voz.** Es una de las aplicaciones de PLN que más éxito ha tenido en la actualidad, ya que es común que las computadoras de hoy tengan esta facilidad. El reconocimiento de voz puede tener dos usos posibles: identificar al usuario o procesar lo que el usuario dicte y ya existen programas comerciales accesibles por los usuarios, por ejemplo: ViaVoice.
- **Recuperación de la información.** Un ejemplo claro de esta aplicación es el siguiente: una persona llega a la computadora y le dice en Lenguaje Natural qué es lo que busca; ésta busca y le dice lo que tiene referente al tema.

En los últimos diez años se han multiplicado las investigaciones para desambiguar palabras automáticamente, crear métodos de identificación y usar las irregularidades encontradas. Los sistemas actuales de recuperación de información en línea carecen de un método inteligente que permita mejorar su eficiencia. Por lo tanto, este trabajo de investigación **se concentra en crear un Método de desambiguación de sentidos de palabras usando grandes recursos léxicos** para ser aplicado en la recuperación de la información y en la navegación en hipertexto.

Objetivos

Objetivo general

Se cree que el disponer de servicios eficientes de recuperación inteligente de información permite mejorar la respuesta a los usuarios que buscan información. En base a esta hipótesis, el objetivo general de esta investigación es **diseñar un nuevo Método de desambiguación de sentidos de palabras usando grandes recursos léxicos** que mejore la pertinencia de la información recuperada.

Objetivos específicos

El desarrollo de métodos para la recuperación de la información es en la actualidad una de las tareas de investigación en Ciencias de la Computación, ya que permite mejorar la eficiencia en la obtención de la información pertinente.

Debido a esto, es estratégico crear métodos que contribuyan a alcanzar esas metas y en consecuencia, es necesario diseñar nuevas técnicas de recuperación inteligente, que permitan la resolución de ambigüedad de sentidos de palabras.

De forma genérica, la WSD consiste en la asociación de una palabra en un texto con una definición o significado dado que la distingue de otros significados atribuibles a esa palabra. La asociación de las palabras a diversos sentidos depende de dos recursos de información: contexto y recursos de conocimientos externos.

El contexto de la palabra a ser desambiguada se considera como el conjunto de palabras que acompaña a la palabra a desambiguar junto con las relaciones sintácticas, categorías semánticas, etc. Los recursos de conocimiento externos son los recursos léxicos (WordNet), enciclopédicos, etc., desarrollados de forma manual o automatizada, que proporcionan datos valiosos para asociar las palabras con los diversos sentidos posibles.

Los **objetivos específicos de este trabajo** se orientan hacia el diseño de un nuevo Método de desambiguación de sentidos de palabras usando grandes recursos léxicos y son:

1. Preparar los recursos léxicos (diccionarios) para usar en la desambiguación.
2. Diseñar un nuevo método de WSD teniendo en cuenta el contexto local del documento donde se ponderen los posibles sentidos en función de esa área limitada dentro del mismo al usar diferentes recursos léxicos (diccionario explicativo Anaya, diccionario WordNet para el español, diccionario de sinónimos) y la semejanza entre las palabras y que permita:
    a. Analizar y determinar los tamaños óptimos del contexto a usar en el proceso de desambiguación.
    b. Calcular los pesos de los sentidos asociados a cada palabra.
    c. Aplicar la intersección de contextos y definiciones.
    d. Aplicar el diccionario de sinónimos.
    e. Limitar de forma automática la sustitución en profundidad.
    f. Aplicar las primitivas semánticas.
3. Preparar una colección de documentos de prueba con ciertos criterios, aplicarle el método y analizar su comportamiento.
4. Investigar el comportamiento del método para el contexto general de todo el documento y realizar el refinamiento del método.
5. Realizar pruebas de eficiencia del método en una colección de documentos en español.

Organización de la tesis

El resto del documento se organiza en antecedentes, método propuesto, resultados, conclusiones, glosario, referencias e índice términos de la siguiente manera.

En el **Capítulo 1, Antecedentes**, se presenta una breve revisión del estado del arte y se mencionan las principales tendencias de investigación en la recuperación de la información, las fuentes de conocimiento, el lenguaje natural y su ambigüedad y, la desambiguación de los sentidos de las palabras. Se introducen los conceptos básicos y se ilustran algunas de sus tareas principales. Se describen y comparan los enfoques de análisis y, se justifica la selección de uno de estos enfoques para este trabajo.

En el **Capítulo 2, Método propuesto**, se presenta el método y cada uno de los pasos necesarios para dar cumplimiento a las metas generales y especificas.

Posteriormente en el **Capítulo 3, Resultados**, se presentan los resultados y se compara el método con otras soluciones.

Finalmente se presentan las **conclusiones** y para ayudar en la comprensión de los resultados de la investigación, se dispone adicionalmente de un **glosario** con definiciones de términos, palabras, siglas, etc. que pueden contribuir al entendimiento del documento. Así mismo se incluyen **referencias,** con las bibliografías consultadas y referenciadas que aparecen enumeradas y ordenadas alfabéticamente. Finalmente, un **índice de términos** con la lista de la ubicación dentro del documento de algunos términos y temáticas ordenadas alfabéticamente para facilitar su rápida localización.

1

Antecedentes

En este capítulo se presenta una breve revisión del estado del arte de la temática y se mencionan las principales tendencias en la investigación, las fuentes de conocimientos, el lenguaje natural y su ambigüedad y, la desambiguación del sentido de las palabras. También se introducen los conceptos básicos y se ilustran algunas de las tareas principales de la WSD, así como la descripción y comparación de los enfoques de los análisis hechos y se justifica la selección de uno de estos enfoques para este trabajo.

Capítulo 1. Antecedentes

En los últimos veinte años, la recuperación de datos ha crecido más allá de las limitaciones sobre textos y de buscar la indización de direcciones útiles para documentos pertenecientes a una colección.

Hoy en día, la investigación en Recuperación de Información (*Information Retrieval* – IR) incluye la modelación, clasificación y categorización del documento, la arquitectura de los sistemas, las interfaces de usuario, la visualización de los datos, el filtrado, los lenguajes, etc.

A pesar de su madurez, hasta hace poco el IR fue visto como un campo de interés estrecho, utilizado principalmente por los bibliotecarios y los expertos de la información. Esta visión tan tendenciosa prevaleció por muchos años a pesar de la difusión rápida entre los usuarios de computadoras personales modernas.

Al principio de los años noventa, un hecho cambió de una vez por todas estas opiniones: la introducción de la World Wide Web. A pesar de tanto éxito, la Web ha introducido nuevos problemas como lo aburrido y difícil que es encontrar la información útil en ella.

El obstáculo principal es la ausencia de un modelo de datos para la Web, lo que implica que la definición y la estructura de la información sean con frecuencia de baja calidad. Estas dificultades han atraído un interés renovado en IR y sus técnicas como posibles soluciones. Consecuentemente, el IR ha ganado un nuevo lugar junto con otras tecnologías. [Baeza,1999].

La recuperación de información con desambiguación del sentido de las palabras es una de las posibles soluciones a la recuperación eficiente de la información. El enfoque dominante para el problema de desambiguación del sentido de las palabras usa los métodos estadísticos, es decir, no se usan algunas fuentes adicionales de conocimiento [Manning y Schütze, 1999]. Por ejemplo, los métodos basados en los clasificadores bayesianos, las redes neuronales, las máquinas de soporte vectorial u otras técnicas del área de la estadística pura.

Por otro lado, los enfoques que usan fuentes adicionales de conocimiento se propusieron hace mucho tiempo [Lesk, 1986], [Hirst, 1987]. La ventaja de estos enfoques basados en conocimiento adicional, es su claridad: se puede seguir el algoritmo paso a paso verificándolo, de tal manera que su decisión es totalmente explícita, no depende de los procesos de aprendizaje y entrenamiento y, en teoría, este enfoque puede llegar al 100% de eficiencia.

Por esa razón, en algunos trabajos recientes se presentaron algunas modificaciones del algoritmo original de Lesk, basadas en el uso de tesauros, diccionarios de sinónimos, diferentes tipos de normalizaciones morfológicas, etcétera. [Wilks, 1998], [Yarowsky, 1992].

En el algoritmo original de Lesk, el sentido de la palabra se representa como un conjunto de cadenas de caracteres que forman la definición, por ejemplo, $banco_1$ = {“*instituto*”, “*financiero*”}. El algoritmo calcula los pesos de los sentidos de las palabras que pertenecen a este conjunto basándose en la intersección del conjunto con los conjuntos de cada sentido de cada palabra y después toma el sentido con el mayor peso.

Desde mediados de la década de los ochenta, pero más en los últimos años, la resolución de la ambigüedad del sentido de las palabras ha sido el tema de investigación de lingüistas, científicos en gestión del conocimiento, inteligencia artificial, sistemas computacionales y lingüística computacional [Mitkov, 1998]. Su importancia radica, entre otras razones, en que:

- Es uno de los fenómenos más complejos dentro del lenguaje natural [Huang, 2000], [Mitkov, 2001]

- Es necesaria en un amplio rango de tareas del Procesamiento del Lenguaje Natural (PLN) como en las interfaces en lenguaje natural, la comprensión del lenguaje, traducción automática, extracción de información y generación automática de resúmenes [Hirst 1981], [Carter 1987], [Fox 1987], [Aone y McKee 1993], [Mitkov, 2001]

La resolución de la ambigüedad del sentido de las palabras es un mecanismo lingüístico para definir el sentido más adecuado de una palabra, según el contexto donde se emplee, definiéndose en función de sus posibles sentidos. Por ejemplo:

(1) El auto se averió, y Nelson tomó el **gato** para levantarlo mientras Hortensia descansaba.
gato: Máquina compuesta de un engranaje de piñón y cremallera que sirve para levantar grandes pesos a poca altura.

(2) El ratón asustó a Hortensia y Nelson llamó al **gato** para que lo cazara.
gato: Mamífero carnívoro, felino, de pelaje suave, que vive en domesticidad y es muy útil en las casas porque persigue a los ratones

(3) Dice Hortensia: “Nelson, eres un **gato**, yo no pude resolver el crucigrama”.
gato: Hombre sagaz y astuto

En el ejemplo puede observarse que la palabra “gato” tiene varios sentidos y según el contexto donde se emplee, significa uno de ellos.

1.1 Recuperación de la información

La Recuperación de Información (*Information Retrieval* – IR) consiste en la tarea de *ordenar* los documentos, tanto de texto como de multimedia, que pertenecen a una *colección* dada de acuerdo a una probabilidad estimada de *relevancia* para las *necesidades* de información del *usuario*. Estas necesidades de información son expresadas generalmente por el usuario en función de las respuestas obtenidas a un requerimiento en un lenguaje no formalizado, por ejemplo sentencias, o un conjunto de términos en un lenguaje natural.

La enorme disponibilidad actual de documentos almacenados electrónicamente, especialmente en plataformas distribuidas, ha transformado la recuperación de información en una disciplina importante. La World Wide Web (WWW) contiene grandes cantidades de información potencialmente interesante y accesible para muchos usuarios.

Las tecnologías actuales de IR (representadas en el contexto Web por "buscadores" o "portales", tales como Altavista, Lycos, Infoseek, etc.) hacen que las necesidades de los usuarios sean satisfechas únicamente de forma parcial o primitiva. El gran número de documentos irrelevantes que usualmente recuperan estos buscadores, y el gran número de documentos relevantes que estos sistemas *no* recuperan, son un obstáculo substancial para una mejor utilización de los recursos de información disponibles en la Web. [Avancini, 2000]

Una de las aproximaciones que ha sido propuesta para mejorar la efectividad en la búsqueda de información es la *categorización* de los documentos, pertenecientes a una colección, según un conjunto predefinido de categorías. La recuperación de información se beneficia con esta técnica porque los documentos se buscan directamente en las categorías relevantes en lugar de buscarlos en la colección completa. Además, la estructura típicamente jerárquica del conjunto de categorías permite que el usuario la recorra y pueda ir *refinando* su búsqueda de forma incremental.

Los motores de búsqueda que emplean esta técnica son, por ejemplo, Yahoo! e Infoseek. La categorización previa de los documentos puede ser una aproximación computacionalmente factible (y aún aplicable) únicamente si ésta se realiza en forma automática. Actualmente, Yahoo! e Infoseek hacen esto de forma manual, empleando catalogadores humanos expertos, lo que tiene un impacto desfavorable en el costo del mantenimiento de estos sistemas.

El acceso a diferentes fuentes de información electrónicas presenta al menos tres complicaciones [Genesereth, 1997]:

- **Distribución**: las consultas no dependen necesariamente de datos provenientes de una sola fuente de información, sino más bien dependen de fragmentos de información que deben ser combinados.
- **Heterogeneidad**: las diferentes fuentes de información generalmente usan diversos lenguajes y protocolos de acceso; por ejemplo, cuando diferentes palabras significan un mismo concepto (computación o informática) y viceversa.
- **Inestabilidad:** las fuentes de información cambian el formato de sus datos o cambian su contenido.

La distribución de las fuentes de información naturalmente implica el empleo de una aproximación distribuida para obtener los datos. El paradigma de agentes [Jennings et al.1998] responde a los requerimientos planteados por estos sistemas, pues facilita las tareas del usuario y del desarrollador [Shoham, 1993]. Los agentes inteligentes pueden localizar, recuperar, clasificar e integrar los documentos.

1.1.1 La información

Cuando se habla de *información* en general podemos encontrar:

- Texto (libros, periódicos, Web, memos, publicaciones, etc.)
- Películas
- Fotos (imágenes en general)
- Televisión, Radio
- Conversaciones telefónicas
- Bases de datos

La información disponible hoy en día es enorme. Como ejemplo, algunas cifras:

Librería del Congreso (*Library of Congress*): ~20 Terabytes

Diálogos: ~9.2 Terabytes

Web: 8+ Terabytes (2006)

Donde:
Gigabyte = 10^9 bytes
Terabyte = 10^12 bytes

Petabyte = 10^15 bytes
Exabyte = 10^18 bytes

De forma general la información se puede jerarquizar (Figura 1) y pasa por distintas etapas en su ciclo de vida(Figura 2).

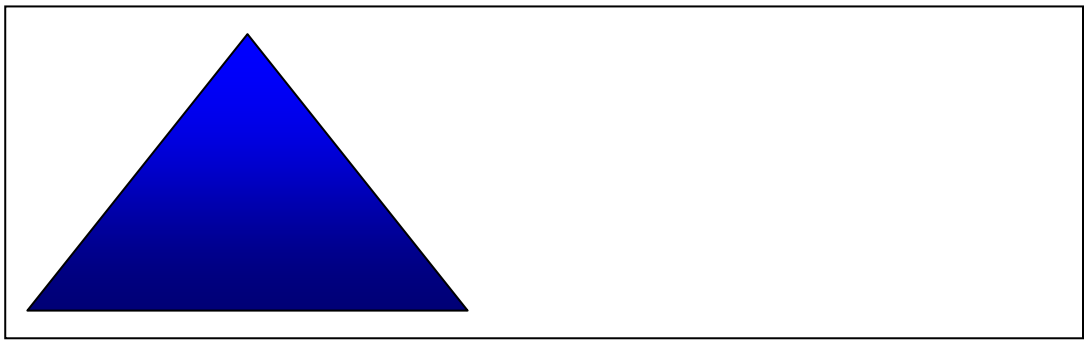


Figura 1. Jerarquía de la información

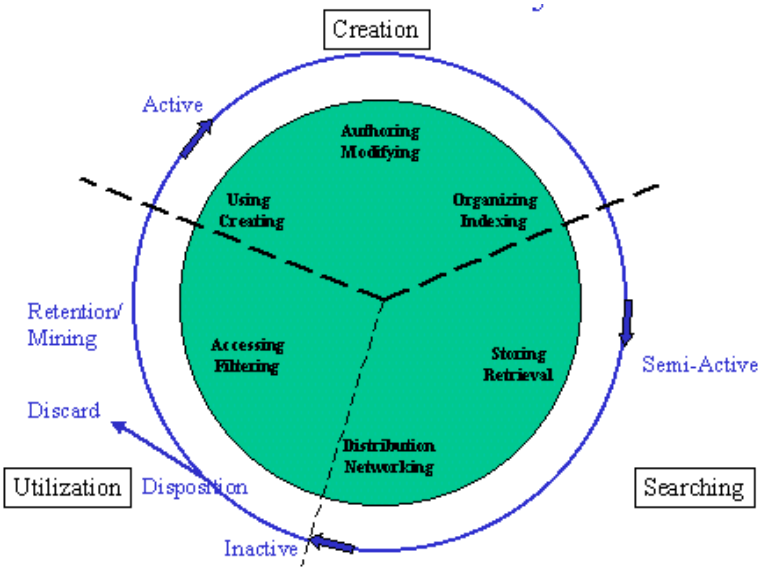


Figura 2. Ciclo de vida de la información

1.1.2 La recuperación

Los pasos para la recuperación van desde las necesidades y la consulta de los usuarios, por las fuentes de información y su procesamiento, hasta la selección y evaluación de los resultados (Figura 3).

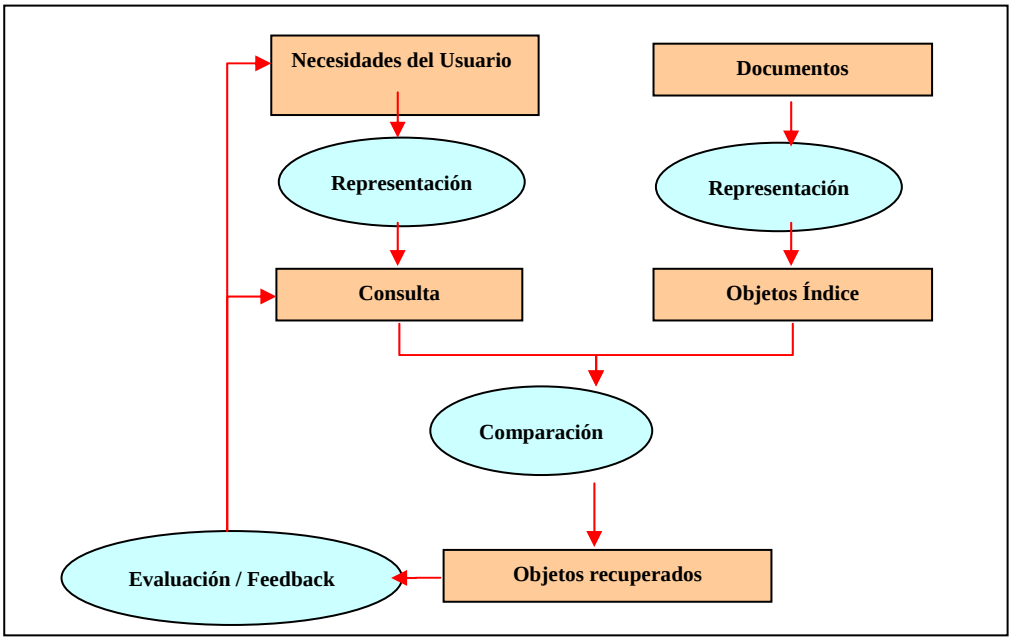


Figura 3. Proceso de recuperación de la información

1.1.3 Antecedentes de los sistemas de recuperación de la información

El interés por los sistemas basados en computadoras para la *recuperación de la información* data de mediados de los cincuenta. Se brinda a continuación una cronología de los sistemas más destacados:

- H.P. Luhn de IBM – recuperación estadística (1958)
- Modelos Probabilísticos de Rand (Maron & Kuhns) (1960)
- Sistemas Booleanos desarrollados en Lockheed - DIALOG (1966)
- Modelo de Espacio de Vectores - SMART (Salton y Cornell 1968)
- Métodos estadísticos y avances teóricos (70s)
- Refinamientos y análisis en aplicaciones y algoritmos (80s)

- Primer sistema de gran escala con recuperación probabilística - *West's legal retrieval system* (1992)
- Avances en bases de datos multimedia (mediados de los noventa hasta el presente)
- Explosión de Internet y la WWW, sistemas de recuperación, integración entre bases de datos y Web (1994 hasta el presente)
- Avances en Bibliotecas Digitales (1995 hasta el presente)
- Interfaces de usuario, sistemas de recuperación de imágenes y vídeo (fines de los 90)

1.1.4 Tendencias de los sistemas de recuperación de la información

Las tendencias en los sistemas de IR apuntan hacia:

- La recuperación de información basada en la lógica.
- Una mejor integración entre procesamiento de lenguaje natural, el aprendizaje por máquinas y las tecnologías de agentes.
- La recuperación de información distribuida, heterogénea y acceso a bases de datos.

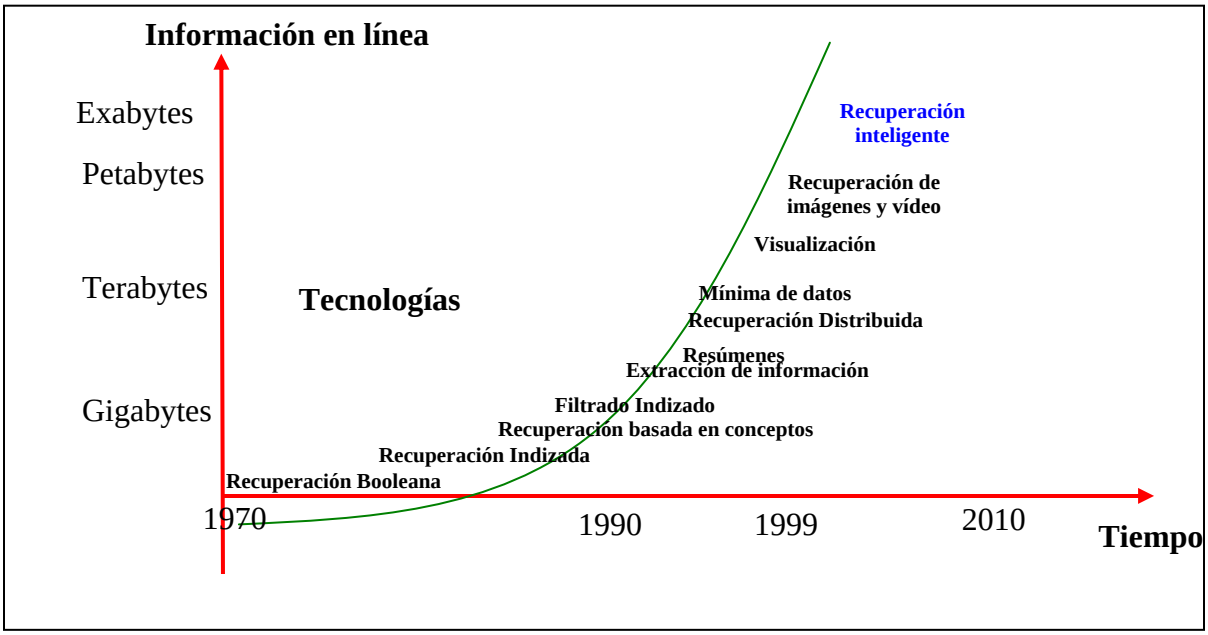


Figura 4. Tendencias de los sistemas de recuperación de la información

1.1.5 Evaluación de los sistemas de recuperación de la información

La evaluación de los sistemas de recuperación de la información abarca generalmente el análisis de grandes sistemas. La evaluación analiza aspectos tales como la asistencia en la formulación de consultas, velocidad en la recuperación, recursos requeridos, presentación de los documentos y habilidad para encontrar documentos relevantes. Estos aspectos se comparan entre los distintos sistemas, donde el factor más usual es la efectividad en la recuperación. [Avancini, 2000]

La *efectividad en la recuperación* depende de la *relevancia* de los documentos recuperados. Esta *efectividad* es usualmente medida en términos de "*recall*" y "*precision*".

Recall, es la porción de material relevante que ha sido recuperada. (igual a Relevantes recuperados / Relevantes en la colección)

Precision, es la porción del material recuperado que es realmente relevante. (igual a Relevantes recuperados / Recuperados)

Es difícil definir precisamente la relevancia de un documento ante una consulta particular porque un documento relevante "juzgado" como útil para satisfacer una consulta depende de: (a) ¿quién hace la consulta?, (b) ¿qué es útil?, (c) ¿en qué contexto fue realizado el juicio?

En las colecciones de documentos reales casi nunca se conoce todo el conjunto de los documentos relevantes. Cualquier modelo de recuperación incluye una definición implícita de relevancia.

Es importante definir qué constituye una "buena" efectividad en un sistema autónomo (como por ejemplo, software que busca documentos y agentes que alertan al usuario), y configurarlo para que logre la máxima efectividad. Además, es importante también poder estimar cómo cambia la efectividad cuando se procesan datos nuevos [Lewis, 1995].

Las *medidas de efectividad* son, como se mencionó anteriormente, *precision* y *recall*; sin embargo, existen otras medidas como el *fallout* que representa la porción de

los documentos que el sistema clasifica en otra clase diferente a la que realmente corresponde.

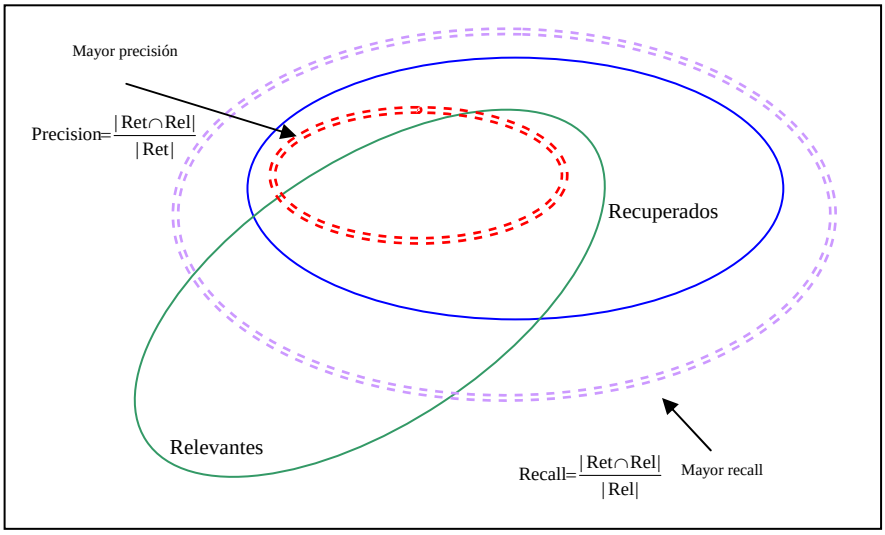


Figura 5. Relación de los documentos relevantes vs recuperados

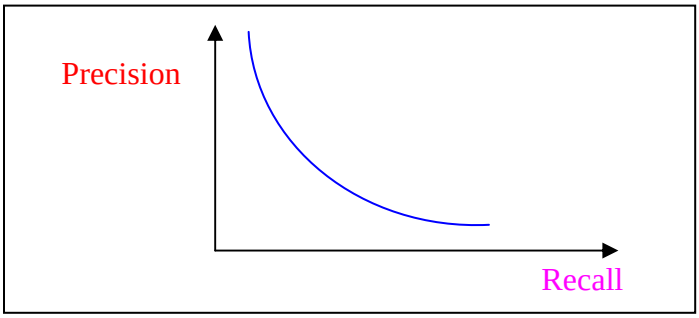


Figura 6. Curva de la relación entre “Precision” y “Recall”

Para *estimar la efectividad* (no siempre es posible calcularla directamente) se puede contrastar la salida del sistema con la de un experto humano. Otra estimación puede ser calculada suponiendo que el juicio de un experto es modelado como una variable aleatoria del tipo *Bernoulli*.

La *efectividad esperada* puede *ser optimizada* si el sistema es capaz de controlar sus propios parámetros. Existen algunos "principios" para hacer esto, por ejemplo, el principio de *ranking* por probabilidad

1.1.6 Empleo de la retroalimentación ("feedback")

El empleo del *feedback* en la recuperación de la información permite mejorar la efectividad de los sistemas. Una sesión con un sistema de recuperación de información por *ranking,* comienza con el usuario que ingresa una consulta textual que describe la información que necesita.

Esta consulta es interpretada por el sistema de IR que intenta separar los documentos relevantes de los no relevantes, estableciendo un *ranking* entre los primeros.

Además, en un sistema de IR que emplea *feedback* el usuario tiene la opción de marcar los documentos del *ranking* que considera más importantes para su necesidad de información. Esta información es proporsionada al sistema que, junto a la consulta original, realiza una nueva recuperación de los documentos relevantes y no relevantes para producir un nuevo *ranking*.

Los resultados experimentales demuestran que los sistemas de IR que usan *feedback son* entre un 20% y un 200% más efectivos que los que solamente usan la consulta inicial [Lewis, 1995a].

1.2 Fuentes adicionales de conocimiento no estadístico

Los recursos lingüísticos son un elemento esencial de la ingeniería lingüística. Constituyen una de las principales formas de representar el conocimiento de la lengua utilizada en los trabajos de análisis que conducen al reconocimiento y la comprensión.

El trabajo de producir y mantener recursos lingüísticos (ej. diccionarios) es una tarea descomunal. De la producción de los recursos se encargan centros de investigación e instituciones públicas con arreglo a formatos y protocolos normalizados que permiten

utilizarlos en muchas de las lenguas. La Asociación Europea de Recursos Lingüísticos (ELRA, "European Language Resources Association") produce buena parte de estos recursos. [HLTTeam, 2002]

1.2.1 Léxicos

Un léxico es un depósito de palabras y de conocimientos sobre ellas. Entre éstos se consideran: la información sobre la estructura gramatical de cada palabra (morfología), la estructura fonética (fonología) o el significado de la palabra en diferentes contextos textuales, por ejemplo, en función de la palabra o del símbolo de puntuación que le antecede. Para ser útil, un léxico debe tener cientos de miles de entradas y son necesarios para cada lengua en particular.

1.2.2 Léxicos especializados

Existen algunos tipos especiales de léxicos que se suelen investigar y producir independientemente de los léxicos de carácter general como se muestran en [HLTTeam, 2002] y que a continuación señalan:

- ***Nombres propios***: Los diccionarios de nombres propios son esenciales para una comprensión eficaz de la lengua, al menos para que éstos puedan ser reconocidos dentro de su contexto como lugares, objetos, personas, o incluso animales. En muchas aplicaciones adquieren un significado especial cuando el nombre es fundamental para la aplicación, por ejemplo, en los sistemas de navegación por voz, en los sistemas de reserva de vacaciones o con la información de los horarios de ferrocarril, que se basan en la gestión automática de las llamadas telefónicas.
- ***Terminología***: En el complejo entorno tecnológico de nuestros días existe una gran multitud de términos que es preciso registrar, estructurar y poner a disposición de las aplicaciones lingüísticas. Muchas de las aplicaciones más productivas de la ingeniería lingüística, como la gestión de documentos técnicos

multilingües y la traducción automática, dependen de que se disponga de los bancos terminológicos apropiados.

- ***Redes de palabras ("wordnets")***: Las redes de palabras describen las relaciones existentes entre las palabras, por ejemplo, los sinónimos, antónimos, sustantivos colectivos, etc. Tienen un gran valor para aplicaciones tales como de búsqueda de información, en herramientas integradas de apoyo a la traducción y en los sistemas ofimáticos inteligentes de creación de documentos.

1.2.3 Diccionario explicativo.

El diccionario explicativo define las palabras por medio de definiciones compuestas por otras palabras, por ejemplo, *banco* es *el instituto financiero*. Parece que en este caso tenemos la relación entre las palabras de definición y la palabra, que está definida, pero no es así, de hecho lo que se define no son las palabras sino los sentidos de las palabras: *banco*$_1$ es *el instituto financiero*, mientras *banco*$_2$ es *el mueble para sentarse*. Sin embargo, en este caso las palabras que se usan en las definiciones son cadenas de caracteres y no los sentidos particulares de las palabras, por ejemplo, en definición de *banco*$_1$ la palabra *instituto* puede significar uno de los siguientes conceptos: una escuela, un centro de investigación, una estructura social o una organización.

Para cualquier aplicación de diccionarios explicativos dentro del área de procesamiento automático de lenguaje natural o de lingüística computacional es necesario eliminar la ambigüedad del sentido de palabras en sus definiciones. [Gelbukh, 1997].

Cuando los grandes diccionarios explicativos existentes, (como el Larousse o el de la Academia Real Española), se usan para facilitar el análisis automático de textos por computadora, presentan los siguientes problemas:

1) las definiciones no son suficientemente estructuradas y formalizadas,

2) se usa un número redundante de palabras para definir a otras palabras en vez de definir todos los términos a través de un número limitado de conceptos básicos como se hace en matemáticas.

En el mundo se dedica un gran esfuerzo a la compilación de los diccionarios semánticos y los tesauros. Como ejemplos de diccionarios semánticos se pueden mencionar los siguientes: el proyecto CYC, quizás el proyecto más grande de este tipo; el diccionario WordNet y la red semántica FACTOTUM SemNet de MICRA, Inc. La mayoría de estos recursos léxicos son para el inglés. El único diccionario semántico para el español, EuroWordNet, está todavía en proceso de desarrollo, tiene una estructura relativamente simple e información relativamente pobre.

Otros ejemplos: el DEX, "*Dictionarul explicativ al limbii române*" (Diccionario explicativo de la lengua romana), Diccionario Explicativo de Vines con palabras del Antiguo y Nuevo Testamento, Diccionario explicativo ANAYA para el español, Diccionario explicativo-combinatorio de la Escuela Semántica de Moscú-Montreal, etcétera.

1.2.4 La cadena de desarrollo y aplicación

El diagrama siguiente (Figura 7) muestra la cadena de actividades de la ingeniería lingüística desde la investigación hasta la entrega de productos y servicios lingüísticos a los usuarios finales. [HLTTeam, 2002]

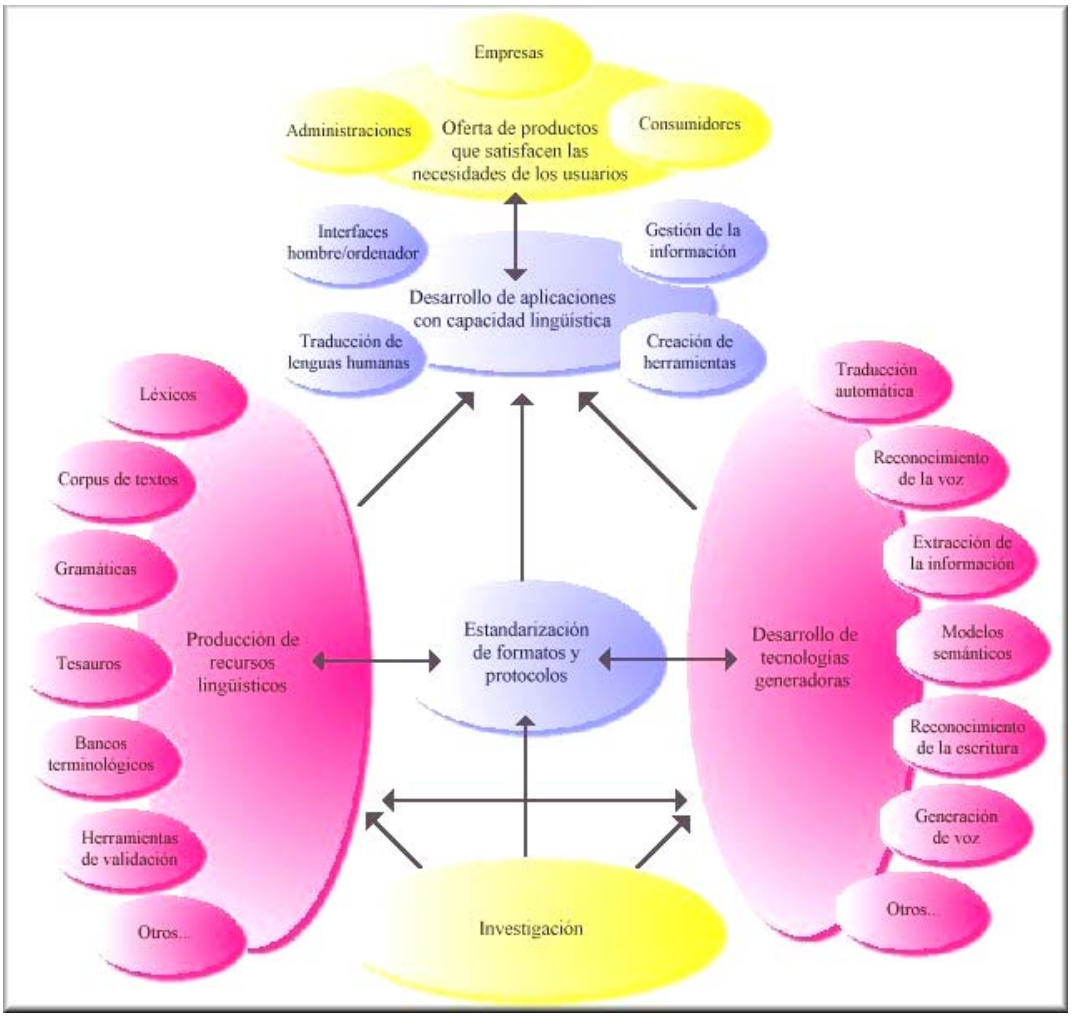


Figura 7. Modelo de las actividades de la Ingeniería Lingüística

El proceso de investigación y desarrollo conduce a la creación de técnicas, la producción de recursos y el establecimiento de normas que constituyen los bloques de construcción básicos.

En términos prácticos, la ingeniería lingüística se aplica en dos niveles.

En el primer nivel, existen diversas clases genéricas de aplicaciones, a saber:

- la traducción de lenguas,
- la gestión de la información (multilingüe),
- la creación de herramientas (multilingües),
- las interfaces persona - máquina (habla y textos multilingües).

En el segundo nivel, estas aplicaciones generadoras se aplican a los problemas del mundo real en todo el espectro social y económico. Así, por ejemplo:

- La gestión de la información puede utilizarse en un servicio de información como base para analizar las solicitudes de información y acoplarlas a una base de datos de texto o imágenes, a fin de seleccionar la información precisa;
- las herramientas de creación se utilizan primordialmente en los sistemas de tratamiento de texto, pero también pueden emplearse para generar textos (por ejemplo, cartas comerciales en idiomas extranjeros) y, en combinación con los sistemas de gestión de la información, para brindar medios de gestión documental;
- la traducción del lenguaje humano se utiliza en la actualidad para ofrecer a los traductores herramientas integradas de apoyo a la traducción y servicios de traducción automática en campos bien delimitados;
- la mayor parte de las aplicaciones pueden disponer de interfaces de usuario en lenguaje natural, incluido el habla, para mejorar su uso.

Por lo general, la capacidad lingüística se añade a los sistemas para incrementar su rendimiento. La Ingeniería Lingüística es, en este sentido, "una tecnología generadora".

1.3 Procesamiento computacional del Lenguaje Natural

1.3.1 El lenguaje

Aunque no existe una definición única, según Lecth, un lenguaje se considera como un conjunto usualmente infinito de oraciones, formado con combinaciones de palabras del diccionario. Es necesario que esas combinaciones sean correctas (con respecto a su sintaxis) y tengan sentido (con respecto a la semántica).

Un lenguaje es la función que expresa pensamientos y comunicaciones entre la gente. Esta función se lleva a cabo por medio de señales y vocales (voz) y posiblemente por signos escritos (escritura).

En este punto podemos distinguir entre dos clases de lenguajes: los naturales (inglés, alemán, español, etc.) y los formales (matemático, lógico, etc.). A continuación damos una breve descripción de estos dos tipos de lenguaje.

1.3.2 El lenguaje natural

Como se mencionó anteriormente, el Lenguaje Natural (LN) es el medio que se utiliza de manera cotidiana para establecer comunicación entre las personas.

Este tipo de lenguaje es el que permite designar las cosas actuales y razonar respecto a ellas; ha sido desarrollado y organizado a partir de la experiencia humana y puede ser utilizado para analizar situaciones altamente complejas y razonar muy sutilmente. La riqueza de sus componentes semánticos da a los lenguajes naturales un gran poder expresivo y valor como una herramienta para el razonamiento sutil. Por otro lado, la sintaxis de un LN no puede ser modelada fácilmente mediante un lenguaje formal, similar a los utilizados en las matemáticas y la lógica. Otra propiedad de los lenguajes naturales es la polisemia, es decir, la posibilidad de que una palabra en una oración tenga diversos significados.

En un primer resumen, los lenguajes naturales se caracterizan por las siguientes propiedades:

1. Han sido desarrollados por enriquecimiento progresivo antes de cualquier intento de formación de una teoría.
2. La importancia de su carácter expresivo se debe fundamentalmente a la riqueza del componente semántico (polisemia).
3. Existe dificultad o imposibilidad de una formalización completa.

1.3.3 El lenguaje lormal

El Lenguaje Formal es aquel que el hombre ha desarrollado para expresar las situaciones que se dan en específico en cada área del conocimiento científico. Las

palabras y oraciones de un lenguaje formal están perfectamente definidas (una palabra mantiene su mismo significado independientemente de su contexto o uso.

Los lenguajes formales están exentos de cualquier componente semántico fuera de sus operadores y relaciones. Los lenguajes formales pueden ser utilizados para modelar una teoría de la mecánica, física, matemática, ingeniería eléctrica, o de otra naturaleza, con la ventaja de que en éstos se elimina toda ambigüedad. [Letch, 1992]

En resumen, las características de los lenguajes formales son las siguientes:

1. Se desarrollan de una teoría preestablecida.
2. Tienen componente semántico mínimo.
3. Posibilidad de incrementar el componente semántico de acuerdo con la teoría a formalizar.
4. La sintaxis produce oraciones no ambiguas.
5. Los números tienen un rol importante.
6. Poseen una completa formalización y por esto, potencialmente posibilitan la construcción computacional.

1.3.4 Procesamiento del lenguaje natural (PLN)

Una meta fundamental de la Inteligencia Artificial (IA) es la manipulación de los lenguajes naturales mediante herramientas de computación. En ello los lenguajes de programación juegan un papel importante ya que forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina.

El PLN es la utilización de un lenguaje natural para la comunicación con la computadora, mediante el entendimiento de las oraciones que le son proporcionadas. El uso de estos lenguajes naturales facilita el desarrollo de programas que realizan tareas relacionadas con el lenguaje, o bien, el de modelos que ayudan a comprender los mecanismos humanos relacionados con el lenguaje.

El uso del lenguaje natural en la comunicación hombre-máquina es a la vez una ventaja y un obstáculo con respecto a otros medios de comunicación. Por un lado es una ventaja, en la medida en que el locutor no tiene que esforzarse para aprender el medio de comunicación a diferencia de otros como los lenguajes de comando o las interfaces gráficas (de 4ta Generación). Su uso es a la vez un obstáculo, porque la computadora tiene una limitada comprensión del lenguaje. Por ejemplo, el usuario no puede hablar con sobrentendidos, ni introducir nuevas palabras, ni construir sentidos derivados, que son tareas que se realizan espontáneamente cuando se utiliza el lenguaje natural.

1.3.5 Niveles del lenguaje

Para continuar el estudio de los lenguajes naturales, es necesario que se conozcan los niveles del lenguaje que se utilizarán para la explicación de la arquitectura de un sistema de PLN. Los niveles de lenguaje que se darán a conocer son los siguientes:

a) Nivel Fonológico: trata de la manera en que las palabras se relacionan con los sonidos que representan.
b) Nivel Morfológico: trata sobre cómo las palabras se construyen a partir de unidades de significado más pequeñas llamadas morfemas, por ejemplo:

   Rápido + Mente == Rápidamente

c) Nivel Sintáctico: trata sobre cómo las palabras pueden unirse para formar oraciones, de acuerdo al papel estructural que cada palabra juega en la oración, y qué sintagmas son parte de otros sintagmas.
d) Nivel Semántico: trata del significado de las palabras y de cómo estos se unen para dar significado a una oración. También se refiere al significado de la palabra independientemente del contexto, es decir de la oración aislada.
e) Nivel Pragmático: trata sobre cómo las oraciones se usan en distintas situaciones y de cómo el uso incide en el significado de las oraciones. Se suele reconocer un subnivel recursivo: discursivo, que trata sobre cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

1.3.6 Arquitectura de un sistema de procesamiento de lenguaje natural

Ahora que ya se conocen los niveles del lenguaje, el siguiente paso es la elaboración de la arquitectura del sistema de Procesamiento del Lenguaje Natural, es decir, cómo va la computadora a interpretar y analizar las oraciones que le sean proporcionadas. A continuación se muestra un esquema de cómo la computadora debe hacer el análisis de estas oraciones.

La explicación de este sistema es sencilla:

a. El usuario le expresa a la computadora qué es lo que desea hacer.
b. La computadora analiza las oraciones proporcionadas en el sentido morfológico y sintáctico, es decir, si las frases contienen palabras compuestas por morfemas y si la estructura de las oraciones es correcta.
c. El siguiente paso es analizar las oraciones semánticamente, es decir, saber cuál es el significado de cada oración y asignar el significado de éstas a expresiones lógicas (cierto o falso).
d. Una vez realizado el paso anterior, se puede hacer el análisis pragmático de la instrucción, es decir, una vez analizadas las oraciones, se analizan todas juntas, tomando en cuenta la situación de cada oración respecto a las oraciones anteriores; así la computadora ya sabe qué es lo que va a hacer, es decir, ya tiene la expresión final.
e. Al tenerse la expresión final, el siguiente paso es su ejecución, para obtener así el resultado y proporcionárselo al usuario.

1.4 Ambigüedad en lenguaje natural

La ambigüedad, en el proceso lingüístico, se presenta cuando pueden admitirse distintas interpretaciones a partir de una representación dada o cuando existe confusión al tener diversas estructuras y no tener los elementos necesarios para eliminar las eventualmente incorrectas. Para desambiguar, es decir, para seleccionar los significados o

las estructuras más adecuadas de un conjunto conocido de posibilidades, se requieren diversas estrategias de solución según el caso. [Galicia-Haro, 1999].

La ambigüedad es el problema más importante en el procesamiento de textos en Lenguaje Natural, por lo que su resolución es la tarea más importante a llevar a cabo y el punto central de esta investigación. Debido a que existe ambigüedad aún para los humanos, su solución se alcanza con lograr la asignación de sentido único por palabra en el análisis de textos, sino eliminando la gran cantidad de variantes que normalmente existen. Con los resultados de este trabajo, se logra diseñar un método para la recuperación de la información basado en la resolución de la ambigüedad del sentido de las palabras para cada texto.

1.5 La desambiguación del sentido de las palabras

1.5.1 Aplicaciones

La desambiguación del sentido de palabras es considerada como uno de los problemas más importantes de investigación en el procesamiento del Lenguaje Natural [Wilks y Stevenson, 1996]. Es esencial para las aplicaciones que requieren la comprensión del lenguaje, como la comprensión de mensajes, la comunicación hombre-máquina, la recuperación de la información y otros. Es requerida, en aplicaciones tales como:

- **La traducción automática.** La desambiguación es esencial para la traducción apropiada de palabras como en el español *banco* que, en dependencia del contexto, puede traducirse como *institución bancaria*, *asiento*, etc. [Weaver, 1949] y [Yngve, 1955].
- **El análisis temático.** Un enfoque común al contenido y al análisis temático es analizar la distribución de categorías predefinidas de palabras con el fin de incluir sólo los sentidos apropiados para el texto [Stone, 1969], [Kelly y Stone, 1975], [Litkowsky, 1997].

- **El análisis gramatical.** La desambiguación en general es útil para el manejo de las restricciones en analizadores [Jensen y Binot, 1987], [Whittemore et al, 1990], [Hindle y Rooth, 1993], [Alshawi y Carter, 1994]. También es útil en la marcación de las partes del habla; detección del género, número, función, etcétera.
- **El procesamiento del habla.** La desambiguación se requiere para la correcta pronunciación en la síntesis del habla [Sproat et al., 1992], [Yarowsky, 1997], y para la segmentación y discriminación de palabras homófonas en el reconocimiento del habla. [Connine, 1990], [Seneff, 1992].
- **El procesamiento de texto.** La desambiguación es necesaria para corregir el deletreo, por ejemplo, para determinar cuándo deben insertarse acentos diacríticos [Yarowsky, 1994] y para la detección y corrección del malapropismo [Hirst, 1998]. Se ha logrado progreso en el área de la representación del conocimiento, con la utilización de redes semánticas aplicadas a la desambiguación.
- **La recuperación de la información y la navegación en hipertexto.** Al realizar búsquedas por palabras claves específicas, se desea eliminar los documentos donde se usa la palabra o palabras en un sentido diferente al buscado; por ejemplo, al buscar referencias judiciales, eliminar documentos que contienen la palabra *court* asociada con *realeza*, en lugar de con *ley* [Salton, 1968], [Salton y McGill, 1983], [Krovetz y Croft, 1992], [Voorhees, 1993], [Schütze y Pedersen, 1995].

En los últimos diez años se han incrementado las investigaciones para desambiguar palabras automáticamente y crear métodos para identificar y usar las irregularidades encontradas. Los sistemas actuales de recuperación en línea carecen de un método de recuperación inteligente que permita mejorar su eficiencia. Por lo tanto, este trabajo de investigación **se concentra en crear un método de recuperación de la información con resolución de la ambigüedad del sentido de las palabras** para ser aplicado en la recuperación de la información y en la navegación en hipertexto.

1.5.2 Elementos de la desambiguación del sentido de las palabras

En general, la desambiguación del sentido de las palabras involucra la asociación de una *palabra* dada en un texto o discurso *con una definición o significado* (sentido)

distinguible de otros significados potencialmente atribuibles a esa palabra. La tarea, por consiguiente, involucra necesariamente dos pasos:

1. La determinación de todos los sentidos diferentes para cada palabra pertinente (por lo menos) en el texto o discurso estudiado.
2. Un medio para asignar cada ocurrencia de una palabra con el sentido apropiado.

En el **Paso 1**, el trabajo más reciente se basa en sentidos predefinidos para la palabra, que incluye: una lista de sentidos como aquellos encontrados en diccionarios cotidianos; un grupo de rasgos, categorías o de palabras asociadas con sinónimos, como en un tesauro.

La definición precisa del "sentido" de una palabra es uno de los debates dentro de la comunidad científica donde no se ve solución en un futuro próximo. Sin embargo, ha habido consenso general en que para palabras homógrafas con diferentes partes del habla (vg. verbo y nombre) la desambiguación puede llevarse a cabo con métodos morfosintácticos [Kelly y Stone, 1975]. Por ello, el trabajo se ha enfocado en distinguir los sentidos entre homógrafos que pertenecen a la misma categoría sintáctica.

En el **Paso 2**, la asignación de sentidos a las palabras, se logra confiando en dos grandes fuentes de información:

- **El contexto de la palabra**, en el sentido amplio; esto incluye la información contenida dentro del texto o discurso en que aparece la palabra, junto con la información extralingüística sobre el texto, como la situación, etcétera.
- **Fuentes de conocimiento externas**, que incluye recursos enciclopédicos léxicos, etc., así como fuentes de conocimiento construidas manualmente que proporcionan datos útiles para asociar las palabras con sus diversos sentidos.

El trabajo de desambiguación reconoce el contexto de la palabra con información de una fuente de conocimiento externa (basado en conocimiento) o información sobre los contextos de casos previamente desambiguados derivados del corpus (basado en corpus). Cualquier método de asociación puede usarse para determinar la mejor selección entre el contexto actual y una de estas fuentes de información para asignar un sentido a cada ocurrencia de la palabra.

1.5.3 El contexto

El contexto es el único medio para identificar el significado de una palabra polisémica (con muchos sentidos). Por consiguiente, el trabajo de WSD se apoya en el contexto de la palabra elegida para proporcionar información que pueda ser usada para su desambiguación. Para métodos basados en datos, el contexto también proporciona el conocimiento previo con el que el contexto actual se compara para lograr la desambiguación. En general, el contexto se determina de dos formas diferentes:

- **El enfoque de "paquete de palabras"**. Aquí, el contexto es considerado como un conjunto de palabras dentro de alguna "ventana" que rodea la palabra designada; es decir, es considerado como un grupo de palabras sin tomar en cuenta sus relaciones con la palabra designada en términos de distancia, relaciones gramaticales, etcétera.
- **Información correlativa**. El contexto es considerado en términos de alguna relación con la palabra designada, como la distancia a ella, las relaciones sintácticas, las preferencias de selección, las propiedades ortográficas, colocación de frases, categorías semánticas, etcétera.

La información del dominio, el contexto del tópico y el contexto local (también conocido como micro contexto) contribuyen a la selección del sentido, pero la importancia de la información y el rol relativo de los diferentes contextos y sus interrelaciones se mantienen en discusión y no han sido bien comprendidos.

1.5.4 El dominio

El uso del dominio para la WSD se evidencia primero en el micro-glosario desarrollado en los trabajos iniciales de traducción automática. La noción de la WSD basada en el dominio está implícita en varios enfoques dentro de la IA (Inteligencia Artificial), como el enfoque de escritura de Schank en el procesamiento de lenguaje natural [Schank y Abelson, 1977] quien agrupó por parejas las *palabras* a los *sentidos* sobre la base del guión (script) activado por el tópico general del discurso.

Este enfoque que activa sólo el sentido de una palabra pertinente al dominio actual del discurso, demuestra sus limitaciones cuando se usa aisladamente. En el famoso ejemplo en inglés "*The lawyer stopped at the bar for a drink*" (El abogado se detuvo en la barra / el bar por una bebida), el correcto sentido de "barra" se asumirá sólo si se apoya en la información de un *guión con relación a la ley*.

Dahlgren en 1988 observó que el dominio no elimina la ambigüedad para algunas palabras; comenta que el nombre común "hand" (mano) tiene al menos 15 sentidos y retiene 10 de ellos en casi cualquier texto. La influencia del dominio depende probablemente de factores como el tipo de texto y la relación entre los sentidos de la palabra designada. [Dahlgren, 1988]

1.5.5 Tópico contextual

El tópico contextual tiene en cuenta palabras individuales que aparecen con un sentido dado de una palabra, normalmente dentro de una ventana de varias oraciones. Diferente al contexto local, se ha utilizado en forma menos consistente. Los métodos que usan el tópico contextual explotan la redundancia en un texto, o sea, el uso repetido de palabras con las que están semánticamente relacionadas a lo largo de un texto sobre un tópico dado. Por ejemplo, la palabra *base* es ambigua, pero su aparición en un documento que contiene palabras como *lanzador*, *bola*, etc. es suficiente para aislar el sentido adecuado para esa palabra (*en el béisbol*).

El uso del tópico contextual se ha discutido en el campo de la recuperación de la información durante varios años [Anthony, 1954] y [Salton, 1968]. Un trabajo reciente en la WSP que ha aprovechado el tópico contextual [Yarowsky, 1992], usa una ventana de 100 palabras para obtener clases de palabras relacionadas y como contexto el texto que rodea a la palabra polisémica considerada, usando el Tesauro de Roget (Roget's Thesaurus).

En [Voorhees et al., 1995] experimentan con varios métodos estadísticos dentro de una ventana de dos oraciones. En [Gale et al., 1993], con una ventana de ±50 palabras y concluyen que las palabras más cercanas dentro del contexto a la buscada contribuyen mejor a la desambiguación, con esto ellos han mejorado sus resultados de 86% a 90% al

extender el contexto de ±6 (típico cuando sólo se considera contexto local) a ±50 palabras alrededor de la palabra designada. En un estudio anterior relacionado [Gale et al., 1992], declara que para un discurso dado, las palabras ambiguas se usan en un solo sentido con una alta probabilidad, es decir, "un sentido por discurso".

El estudio de [Yarowsky, 1993] indica que mientras la información dentro de una ventana grande puede ser usada para desambiguar nombres, para los verbos y adjetivos, el tamaño de la ventana utilizable se reduce dramáticamente con la distancia a la palabra buscada. Esto apoya la idea de que para la desambiguación se requiere tanto del contexto local como el del tópico, y apunta hacia la cada vez más aceptada noción de que para diferentes clases de palabras se pudieran necesitar diferentes métodos de desambiguación.

En [Leacock et al., 1998] se cambia el enfoque al combinar el contexto del tópico y con el contexto local, lo que demuestra que ambos se requieren para lograr resultados consistentes en las palabras polisémicas de un texto [Towell y Voorhees, 1998]. Así, consideran el papel del contexto local vs tópico, e intentan evaluar la contribución de cada uno. Sus resultados indican que *para un clasificador estadístico, el contexto local es superior al contexto de tópico como un indicador del sentido*. Sin embargo, no está claro aún si tal distinción es significativa en el trabajo de la WSD. Puede ser más útil considerar los dos dentro de un "continuum" y considerar el papel e importancia de la información contextual como una función de la distancia a la palabra considerada.

1.5.6 Contexto local

El contexto local ha sido el más utilizado en la mayoría de los trabajos de WSD. El contexto local utiliza la ocurrencia de una palabra como una fuente primaria de información para la WSD. El contexto local o micro-contexto es generalmente considerado como **una ventana** pequeña de palabras que rodea la ocurrencia de una palabra en un texto o discurso; esto va desde unas pocas palabras de contexto hasta la frase completa en la que aparece la palabra designada para la desambiguación.

Aquí, el contexto es considerado como todas las palabras o grupo que caen dentro de alguna ventana que rodea a la palabra designada, sin contemplar relaciones de

distancia, sintácticas, u otras. En general, el enfoque de paquete de palabras ha mostrado funcionar mejor para los nombres que para los verbos, pero es menos eficaz que los métodos en el que se toman en cuenta otras relaciones. Sin embargo, como demostró el trabajo de [Yarowsky 1992], este enfoque es más "barato" que aquellos que requieren un proceso más complejo y logra suficiente desambiguación para algunas aplicaciones. Por ser el contexto local uno de los más utilizado, vale la pena examinar algunos de sus parámetros.

1.5.6.1 Distancia

Yarowsky [Yarowsky 1993, 1994] examinó diferentes tamaños de ventanas de contexto local e incluyó contextos de 1 a k palabras, así cómo pares de palabras con desplazamientos –1 y –2, –1 y +1, y +1 y +2, y los ordenó mediante una relación logarítmica para encontrar la evidencia más fiable para la desambiguación. Yarowsky hace la observación de que el valor óptimo de **k** varía con el tipo de ambigüedad, y sugiere que *las ambigüedades locales necesitan sólo una ventana k de 3* ó *4 palabras*, mientras que *las ambigüedades de tópico* o basadas en la semántica requieren una ventana mayor de 20-50 palabras. [Leacock et al., 1998], usaron una ventana local de ±3 palabras de clase abierta y confirmaron en sus pruebas que dicho número mostró la mejor funcionalidad.

1.5.6.2 Colocación

El término "colocación" fue popularizado por Firth en su trabajo titulado **Modos del Significado**: "*Uno de los significados de **asno** es su colocación habitual con el precedente inmediato **eres tonto**...*" [Firth, 1957]. Él enfatiza que esa colocación no es ninguna simple co-ocurrencia, sino que es la '*habitual*' o '*usual*'. Posteriormente se hicieron varios intentos para definir con más precisión el término en el marco de la teoría lingüística moderna [Halliday, 1961, 1966], [Haas, 1966], [Lyons, 1966], [McIntosh, 1966], [Sinclair, 1966], [Van Buren, 1967].

La definición planteada en [Halliday, 1961] es la más manejable en términos computacionales: "*...es la asociación sintagmática de elementos léxicos, textualmente*

cuantificable, como la probabilidad que ocurrirá a ***n*** *remociones (una distancia de* ***n*** *elementos léxicos) de un elemento* ***x****, los elementos* ***a, b, c****...*". Basado en esta definición, una colocación significativa puede definirse como "*una asociación sintagmática entre los elementos léxicos, donde la probabilidad de co-ocurrencia del elemento* ***x*** *con los elementos* ***a, b, c,****...es mayor que la esperada*" [Berry-Rogghe, 1973]. Es en este sentido en el que la mayoría de los investigadores en WSD utilizan el término.

Yarowsky [Yarowsky, 1993] utiliza explícitamente el uso de las colocaciones en la WSD, pero reconoce la adaptación de la definición a su propósito como "la co-ocurrencia de dos palabras *en alguna relación definida*". Él examina una variedad de relaciones de distancia, pero también considera la adyacencia para cada palabra del discurso (por ejemplo, primer nombre a la izquierda). Determinó que en casos de ambigüedad binaria existe "un sentido por colocación", es decir, en una colocación dada a una palabra se usa en un solo sentido con un 90-99% de probabilidad.

1.5.6.3 Relaciones sintácticas

Earl [Earl, 1973] utilizó exclusivamente la sintaxis para la desambiguación de las palabras en la traducción automática. En la mayoría de los trabajos de WSD usa la información sintáctica junto con alguna otra información. El uso intensivo de restricciones ponderadas de selección en el trabajo de IA descansa en el análisis completo, marcos, redes semánticas, etc. [Hayes, 1977a , 1977b], [Wilks, 1973, 1975] y [Hirst, 1987].

En otros trabajos, la sintaxis se combina con información de la frecuencia de colocación. Es decir, combina la información de la *colocación* con reglas para determinar, por ejemplo, la presencia o ausencia de determinadores, pronombres, complementos nominales, así como preposiciones y relaciones del sujeto-verbo, del verbo-objeto, etc., [Kelly y Stone, 1975], [Atkins, 1987] y [Dahlgren, 1988].

En fechas un poco más recientes, los investigadores han evitado el procesamiento complejo usando un análisis ligero o parcial. En su trabajo de desambiguación con *nombres*, Hearst [Hearst, 1991] segmenta el texto en *nombres* y *frases preposicionales* y *grupos verbales* y desecha toda la información sintáctica adicional. Examina elementos

que están dentro de ± 3 segmentos de la frase considerada y combina la evidencia sintáctica con otros tipos de evidencia, como el uso de mayúsculas.

Yarowsky [Yarowsky, 1993] determinó varios comportamientos basados en la categoría sintáctica, por ejemplo, que los *verbos* generan más información de desambiguación de sus *objetos* que de sus *sujetos*, que los *adjetivos* generan casi toda la información de desambiguación de los *nombres* que modifican, y los *nombres* son mejor desambiguados por los *adjetivos* o *nombres* directamente adyacentes. En trabajos más recientes la información sintáctica es, a menudo, la parte de discurso usada invariablemente junto con otros tipos de información [McRoy, 1992], [Bruce y Wiebe, 1994], [Leacock et al., 1998]. La evidencia sugiere que se necesitan diferentes tipos de procedimientos de desambiguación dependiendo de la categoría sintáctica y de las características de la palabra designada [Yarowsky, 1993], [Leacock, et al., 1998].

Muy pocos estudios han investigado los tres tipos de enfoque para la desambiguación, los más reciente utilizan sólo contexto local. Ésta es otra área donde se requiere el estudio sistemático para la WSD.

1.5.7 Los sentidos

Aunque hay alguna validez psicológica de la noción del sentido de las palabras [Simpson, Burguess, 1988] y [Jorgensen, 1990] los lexicógrafos por sí mismos están conscientes de la falta de acuerdo en los sentidos y sus divisiones [Malakhovski, 1987], [Robins, 1987], [Ayto, 1983], [Stock, 1983]. El problema de la división del sentido ha sido objeto de discusión desde la antigüedad: Aristóteles consagró una sección de sus *Temas* a este asunto en el 350 AC. Desde entonces, filósofos y lingüistas han continuado discutiendo el tema ampliamente [Quine, 1960], [Asprejan, 1974], [Lyons, 1977], [Weinreich, 1980] y [Cruse, 1986], pero la falta de solución se mantiene por más de 2,000 años.

La idea de Aristóteles de que las palabras corresponden a los objetos específicos y conceptos ha sido desplazada en el siglo XX por las ideas de Saussure y otros [Hjemslev, 1953], [Nida, 1966]. Por ejemplo, Antoine Meillet: "*el sentido de una palabra sólo es definido por el promedio de sus usos lingüísticos*" [Meillet, 1926]. En otras palabras no

hay ningún sentido, sino sólo usos: *"No busque el significado, sino el uso";* son puntos de vista similares que se encuentran en las teorías de significado [Bloomfield, 1933] y [Harris, 1954], donde el significado es una función de la distribución y de la situación semántica y el sentido o los sentidos de una palabra son vistos como una abstracción del rol que juega sistemáticamente en el discurso.

El proyecto COBUILD [Sinclair, 1987] adopta esta visión del significado al intentar fijar los sentidos del diccionario en el uso actual mediante divisiones del sentido basándose en los agrupamientos de citas en un corpus. [Atkins, 1987], [Fillmore y Atkins, 1991], [Kilgarriff, 1998] adoptan también implícitamente el punto de vista de [Harris, 1954], donde cada distinción del sentido se refleja en un contexto distinto. Una visión similar se observa en los métodos basados en clases [Brown et al., 1992], [Pereira y Tishby, 1992], [Pereira et al., 1993]. En [Manning y Schütze, 1999] se continúa por este camino y se propone una técnica que evita el problema de distinción del sentido total: *crear agrupamientos del sentido de un corpus en lugar de confiar en una lista de sentidos pre-establecidos*.

1.5.8 Enfoques de los análisis

En cierto sentido el trabajo de WSD ha regresado, recientemente, a los métodos empíricos y a los análisis basados en corpus que caracterizan algunos de los esfuerzos iniciales para resolver el problema. Con mayores recursos y métodos estadísticos mejorados a su disposición, los investigadores están mejorando los resultados de los pioneros, pero parece que se ha llegado al límite de lo que puede lograrse en el marco actual con técnicas y estructuras de representación que impiden distinguir entre lexicones, bases de conocimiento y modelos estadísticos de corpus de texto para el PLN [Dolan et al., 2000].

Por supuesto, la WSD es en parte problemática debido a la dificultad inherente para determinar o incluso definir el sentido de la palabra y esto parece que no será fácilmente resuelto en el futuro cercano [Ravin y Leacock, 2000]. No obstante, parece claro que la investigación actual en la WSD podría beneficiarse considerando las teorías

del significado, el trabajo en el área léxico semántica y el uso de grandes fuentes de conocimientos.

De los enfoques de análisis para la resolución de la ambigüedad del sentido de las palabras son dos los que más influencia han tenido en el área:

- Mediante métodos estadísticos.
- Mediante fuentes adicionales de conocimiento.

1.5.8.1 Enfoque mediante métodos estadísticos

En este enfoque no se usan algunas fuentes adicionales de conocimiento para la resolución de la ambigüedad [Manning y Shutze, 1999] y que tradicionalmente se apoya en los corpus.

Un corpus es una muestra amplia de la lengua escrita o hablada que proporciona las bases para:

- analizar la lengua y determinar sus características;
- entrenar a las máquinas, para adaptar su comportamiento a circunstancias específicas;
- verificar empíricamente una teoría lingüística;
- ensayar una técnica o aplicación de ingeniería lingüística a fin de determinar su buen funcionamiento en la práctica.

Existen corpus nacionales que contienen cientos de millones de palabras, pero se trata de corpus construidos para fines concretos. Por ejemplo, un corpus puede contener grabaciones de conductores de automóvil para simular un sistema de control capaz de reconocer órdenes verbales y con ello determinar las necesidades de los usuarios con vistas a su producción comercial.

Como ejemplo de métodos usando enfoques estadísticos, se tienen los algoritmos que se basan en los clasificadores bayesianos, las redes neuronales, las máquinas vectoriales de apoyo u otras técnicas de la estadística pura . A continuación se presentan

los tres modelos de recuperación de información más relevantes basados en las redes bayecianas .

El primero, denominado *Inference Network Model* (Modelo de las redes de interferencias), fue desarrollado por Croft y Turtle, está constituido como una red bayesiana en la que se distinguen a su vez dos subredes: la red de documentos que es fija para una colección dada y con dos tipos de nodos: término y documento (de los nodos "documento" salen arcos hacia los nodos "término" por los que han sido indizados), y la red de la consulta, que se crea cuando el usuario propone una consulta al Sistema de Recuperación de Información (SRI) y, contiene nodos "consulta" y nodos "término" (los arcos van de los nodos "término" a los nodo "consulta"). [Turtle y Croft, 1990]

Ambas subredes se conectan por medio de los nodos término que existen en ambas desde los nodos de la red de documentos a la de consultas. Una vez que se han estimado las probabilidades, la inferencia se hace a instancias de cada documento sucesivamente y calculando la probabilidad de que la consulta quede satisfecha dado el documento que ha sido observado, es decir, p (Q I d). Una vez que todas las propagaciones hayan finalizado, se genera el correspondiente ordenamiento de documentos.

Estrechamente relacionado con este trabajo está el método conocido como Ghazfan que presenta un modelo básicamente igual al anterior, pero con la diferencia de que cambia la orientación de los arcos. Formalmente, para una consulta Q, los documentos se ordenan según la probabilidad p (d I Q). Para ello se instancian los nodos de la consulta propagando una vez y calculando así la probabilidad de que cada documento sea relevante para la consulta dada. [Ghazfan, 1996]

Por último, Ribeiro [Ribeiro, 1996] presenta el modelo llamado *Belief Network Model* (Modelo de las redes de creencia), donde se consideran únicamente dos tipos de nodos, documentos y términos, enlazados por arcos de los segundos a los primeros. En este modelo, la consulta se considera como un tipo especial de documento, se propaga también una única vez (como Ghazfan) y se obtiene el ordenamiento según p (d I Q).

Los tres modelos mencionados hacen suposiciones de independencia entre términos, y por tanto, no establecen arcos directos entre nodos término. Los modelos *Inference Network* y *Belief Network* no aplican ningún algoritmo de propagación como tal, sino que debido a la topología que tienen sus grafos pueden evaluar la probabilidad de manera directa, con resultados análogos a los de la propagación. [Campos, 2001].

1.5.8.2 Enfoque usando las fuentes adicionales de conocimiento

En este enfoque se hace uso de los grandes recursos lingüísticos para la resolución de la ambigüedad tales como los tesauros, los diccionarios de sinónimos, los diferentes tipos de normalizaciones morfológicas, etcétera.

Durante los últimos años se han realizado algunas investigaciones sobre WSD basadas en el conocimiento. Lesk [Lesk, 1986] propone un método para descifrar el sentido de una palabra en un contexto según el número de coincidencias que aparecen entre el contexto y la definición del diccionario explicativo.

Cowie [Cowie, 1992] describe un método para resolver la ambigüedad léxica de textos basado en la definición dada en *Longman's Dictionary of Contemporary English* (Diccionario Inglés Contemporáneo de Longman) (LDOCE, por sus siglas en inglés) con el que obtiene el 47% en cuanto a distinguir los sentidos y un 72% para las palabras homógrafas.

En [Yarowsky, 1992] se derivan clases de palabras a partir de palabras en categorías comunes del *Roget's Internacional Thesaurus*. En [Wilks, et al. 1993] se utilizan las co-ocurrencias de datos extraídos del LDOCE para construir vectores de contexto y de sentidos asociados a las palabras. En [Voorhees, 1993] se define la construcción denominada *hood*, que utiliza los hipónimos para nombres incorporados en WordNet.

En [Sussna, 1993] se define una métrica basada en la distancia semántica entre los términos de un texto, que consiste en asignar pesos a los enlaces de WordNet según los tipos de relación (sinónimos, hiperónimos, etc.) en el conteo del número de arcos del mismo tipo que salen del nodo y en la profundidad total del arco.

En [Resnik, 1995] se define una métrica basaba en la similitud semántica para las palabras en la jerarquía WordNet. En [Aguirre-Rigau, 1996] se combina un conjunto de algoritmos no supervisados para desambiguar nombres del corpus Semcor. En [Rigau-Aguirre, 1997] se combina un conjunto de algoritmos no supervisados para desambiguar el sentido de las palabras en un corpus no etiquetado.

En [Hale, 1997] se presentan los resultados obtenidos de la combinación de *Roget's Internacional Thesaurus* y la taxonomía de WordNet con la similitud semántica como medida. En [Stetina y Naga 1998] se introduce un método para la WSD, basado en un corpus de entrenamiento etiquetado sintáctica y semánticamente. Este método explota la información del contexto de la oración y sus relaciones semánticas.

En [Resnik, 1999] se presenta una medida para la semejanza semántica presente en una taxonomía IS-A y la aplica en un algoritmo para resolver las ambigüedades sintácticas y semánticas. En [Mihalcea y Moldovan 1999] se expone un método para desambiguar nombres, verbos, adverbios y adjetivos de un texto sobre la base de la referencia del sentido proporcionado por WordNet. En [Montoyo, 2001] se presenta un método que resuelve la ambigüedad léxica de nombres en textos escritos en inglés basado en la taxonomía de nombres que utiliza WordNet.

1.5.9 Enfoque utilizado en este trabajo

Al valorar los dos enfoques analizados, el enfoque mediante métodos estadísticos y el enfoque mediante fuentes de conocimiento, se decidió usar el segundo enfoque.

Las ventajas del enfoque basado en fuentes adicionales de conocimiento son:

- Su claridad: se puede verificar el algoritmo paso por paso.

- Su decisión es totalmente explícita.
- En teoría este enfoque puede alcanzar el 100% de eficiencia.
- No depende de procesos de aprendizaje y de entrenamiento.

2

Método propuesto

En este capítulo se presenta el Método de desambiguación de sentidos de palabras para el idioma español y cada uno de los pasos necesarios para dar cumplimiento a las metas generales y específicas.

Capítulo 2. Método propuesto

2.1 Definición del problema

Uno de los problemas más serios de la recuperación de la información en los portales de Internet (por ejemplo, los portales dinámicos Altavista, Google, Yahoo!, etc.) y en las bibliotecas digitales (por ejemplo, la Biblioteca del Congreso de los E.E.U.U.), es las diversas respuestas con muy baja pertinencia que brindan con respecto a los intereses del usuario.

Por ejemplo, un economista busca *"historia del banco"* y obtiene respuestas sobre los *"bancos de arena"*, *"bancos de madera"* e *"instituciones financieras"*. Un músico busca *"formato de letra"* y obtiene respuestas sobre el *"documento comercial de pago"*, *"letras del alfabeto"* y *"letras musicales"*. Estas imprecisiones se deben a los distintos sentidos que tienen estas palabras.

La desambiguación del sentido de las palabras, WSD por sus siglas en inglés (*Word Sense Disambiguation*), se considera uno de los problemas de investigación con mayor importancia en el campo del procesamiento del lenguaje natural. Es esencial para las aplicaciones que requieren la comprensión del lenguaje, tales como la comprensión de los mensajes, la comunicación hombre-máquina y la recuperación de la información.

En los últimos diez años han crecido en número las investigaciones para desambiguar las palabras automáticamente y crear métodos para identificar y usar las irregularidades encontradas. Los sistemas actuales de recuperación de información en línea carecen de un método de recuperación inteligente que permita mejorar su eficiencia. Por lo tanto, este trabajo de investigación **se concentra en crear un Método de desambiguación de sentidos de palabras para el idioma español,** el cual será aplicado en la recuperación de la información en textos y en la navegación en hipertexto.

2.2 Objetivo general.

Se cree que el disponer de eficientes servicios de recuperación inteligente de la información se podría mejorar la calidad de la respuesta a los usuarios que buscan información. A partir de esta suposición, el objetivo general de nuestra investigación es **diseñar un nuevo Método de desambiguación de sentidos de palabras para el idioma español** que mejore la pertinencia de la información recuperada.

2.3 Objetivos específicos.

El método a desarrollar se basa en el análisis de las palabras que se encuentran en el contexto de la palabra en cuestión. La idea básica, es que las palabras que se encuentran en el contexto cercano se repiten en la definición del sentido de la palabra.

Para detectar la repetición se usarán recursos léxicos de diferentes tipos, teniendo en cuenta no sólo la repetición textual, sino también las relaciones en los diccionarios (sinónimos, entre otros.)

El método a emplear requiere el diseño y la implantación de algoritmos basados en el uso de los recursos léxicos (diccionarios de diferentes tipos: diccionario explicativo Anaya, diccionario WordNet para español, diccionario de sinónimos, representación de las palabras a través de las primitivas semánticas). El método propuesto aplica estos algoritmos y toma la decisión sobre el sentido correcto, calculando los pesos del resultado para cada uno de los sentidos de la palabra.

Para lograr el objetivo general del trabajo es necesario alcanzar los siguientes objetivos específicos:

1. Preparar los recursos léxicos que se usarán en la desambiguación.
2. Diseñar el método de recuperación de la información con desambiguación del sentido de las palabras teniendo en cuenta el contexto local del documento, mediante la ponderación de los posibles sentidos en función de áreas limitadas dentro del documento brindadas por diferentes recursos léxicos (diccionario explicativo Anaya, diccionario de sinónimos Océano y WordNet para español).
3. Preparar una colección de documentos para hacer pruebas de funcionamiento del método propuesto según ciertos criterios.
4. Analizar el funcionamiento del Método de desambiguación de sentidos de palabras y determinar el tamaño óptimo del contexto a usar para la desambiguación.
5. Realizar pruebas de eficiencia del método propuesto con una colección de documentos en español usando un prototipo.

2.3 Metas particulares desarrolladas para la solución de los objetivos.

Para el desarrollo del método y el logro de los objetivos específicos, el trabajo se centró en la solución de los siguientes elementos:

1. Preparación de los recursos léxicos a utilizar en la desambiguación.

1.1. Adaptación del bloque de análisis morfológico del español para la resolución de la ambigüedad del sentido de las palabras en los recursos léxicos.

1.2. Aplicación del análisis morfológico para normalizar las palabras de las definiciones del diccionario explicativo.

1.3. Desarrollar una versión simplificada del método de desambiguación de los sentidos de las palabras en las acepciones del Diccionario Explicativo ANAYA.

1.4. Desarrollar una versión simplificada del método de desambiguación para escoger los sentidos de los sinónimos en el diccionario de sinónimos Océano y WordNet.

1.5. Convertir el archivo de texto WordNet para español en un formato de base de datos.

1.6. Preparar la representación de las palabras como primitivas semánticas en un formato de base de datos.

2. Diseñar el Método de desambiguación de sentidos de palabras.

2.1. Diseñar el método de sustitución de las palabras por sus definiciones según la profundidad y definir una profundidad óptima.

2.2. Diseñar el método de ponderación de los diferentes recursos léxicos: diccionario explicativo ANAYA y los diccionarios de sinónimos de Océano y WordNet.

3. Preparación de la colección de documentos para hacer pruebas de funcionamiento del método desarrollado.

3.1. Definir los criterios de representatividad de los textos.

3.2. Recopilar la colección de documentos de Internet y los corpus que existen en el Laboratorio de Lenguaje Natural del CIC-IPN.

3.3. Realizar el marcado manual de algunas partes de los textos para los experimentos.

4. Análisis del funcionamiento del método y determinación del tamaño óptimo del contexto a usar para la desambiguación.

5. Realización de las pruebas de eficiencia del método propuesto en una colección de documentos en español usando el prototipo.

5.1. Escoger dos métodos como "línea de base" (método original de Lesk, método modificado de Lesk).

5.2. Desarrollar un prototipo y realizar pruebas y mediciones (de tiempo, aciertos y ambigüedades).

5.3. Comparar los resultados con otras soluciones anteriores.

2.4 Explicación detallada del trabajo realizado para cada una de las metas de la investigación.

1. Preparación de los recursos léxicos que se usan en la desambiguación.

1.1. Adaptación el bloque de análisis morfológico de español para la resolución de la ambigüedad del sentido de las palabras en los recursos léxicos.

- Se diseñó la función de utilización de la base de datos morfológica que existe en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC para realizar el análisis morfológico y la normalización.

1.2. Aplicación el análisis morfológico para normalizar las definiciones del diccionario explicativo.

- Se diseñó la función de utilización de la base de datos del Diccionario Explicativo ANAYA que existe en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC para realizar el análisis morfológico y la normalización.
- Se resolvió la homonimia presente en las partes de la oración (o del párrafo) de las definiciones con dos variantes:
    - Mediante el desarrollo de las heurísticas sintácticas. Ver Anexo 2. Algoritmo de desambiguación morfológica
    - Mediante el análisis sintáctico parcial de las relaciones entre las palabras usando el software *parser sintáctico* que existe en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC-IPN. Es parcial por la posibilidad de que falle el acomplamiento sintáctico para todo el texto.
- Se eliminaron las palabras auxiliares tales como; las preposiciones, los verbos auxiliares, los artículos y pronombres y se dejaron los verbos, los adjetivos, los sustantivos y los adverbios.
- Se normalizaron las palabras que conforman las definiciones del Diccionario Explicativo ANAYA, mediante la sustitución de todas las palabras que aparecen en las definiciones por sus *lemas*. Ver Anexo 1. Algoritmo de normalización morfológica.

1.3. Pasos del desarrollo de la versión simplificada del método de desambiguación para desambiguar el sentido de las palabras en las definiciones del Diccionario Explicativo ANAYA.

- Seleccionar la palabra a analizar (a).
- Seleccionar las definiciones de la palabra a analizar (A)
- Escoger para definición A ($P_{A1...n}$), el sentido correcto:
    - Tomar la definición ($B_{1...n}$) de cada palabra $\mathbf{P_{A1...n}}$ que aparece en la definición A
    - Determinar el grado de semejanza entre **A** y $\mathbf{B_{1...n}}$ con un valor normalizado entre 0 y 1:
        - Determinar el aporte de la palabra **a** en cada definición $\mathbf{B_{1...n}}$
        - Determinar el aporte de los sinónimos cercanos de la palabra **a** en cada definición $\mathbf{B_{1...n}}$
        - Determinar el aporte de los sinónimos lejanos de la palabra **a** en cada definición $\mathbf{B_{1...n}}$
        - Determinar el aporte de los sinónimos cercanos de las palabras de la definición **B** $\mathbf{(P_{B1...m})}$ en la definición **A**.
        - Determinar el aporte de los sinónimos lejanos de las palabras de la definición **B** ($\mathbf{P_{B1...m}}$) en la definición **A**.
        - Determinar la profundidad recursiva en un nivel x con un valor entre **0** y **P**, limitado por la relación entre los aportes recursivos que se consideren significativos.
        - Determinar el aporte recursivo del grado de semejanza entre $\mathbf{B_{1...n}}$ y las definiciones de cada palabra de $\mathbf{P_{B1...m}}$ $\mathbf{(C_{1...m})}$ con un valor normalizado entre **0** y **1,** aplicando nuevamente el mismo algoritmo que se aplicó a la semejanza entre **A** y $\mathbf{B_{1...n}}$, deteniendo la profundidad recursiva en el nivel **x.**
    - Escoger como sentido correcto de cada palabra de la definición A ($P_{A1...n}$), el que mayor grado de semejanza obtenga.
- Redefinir la definición **A** en función de los sentidos correctos para cada una de las

palabras que la integran **($P_{A1...n}$).**

1.4. Pasos del desarrollo de la versión simplificada del método de desambiguación para escoger los sentidos de los sinónimos en el diccionario de sinónimos Océano y en WordNet.

- Seleccionar la palabra a analizar (**a**).
- Seleccionar la definición de la palabra a analizar (**A**)
- Seleccionar los sinónimos de **a ($b_{1..n}$)**.
- Escoger para cada sinónimo de (**a**) (**$b_{1...n}$**), el sentido correcto:
    - Seleccionar la definición de cada palabra **$b_{1...n}$** que aparece (**$B_{1...n}$**)
    - Determinar el grado de semejanza entre **A** y **$B_{1...n}$** con un valor normalizado entre **0** y **1**:
        - Determinar el aporte de la palabra (**a**) en cada definición **$B_{1...n}$**
        - Determinar el aporte de las palabras de la definición **$P_{A1..m}$** de (**a**) en cada definición **$B_{1...n}$**
        - Determinar el aporte de los sinónimos cercanos de las palabras de la definición **B** (**$P_{B1...m}$**) en la definición **A**.
        - Determinar el aporte de los sinónimos lejanos de las palabras de la definición **B** (**$P_{B1...m}$**) en la definición **A**.
    - Escoger como sentido correcto de cada sinónimo (**$b_{1...n}$**) de la palabra (**a**), el que mayor grado de semejanza obtenga.
- Redefinir los sinónimos (**$b_{1..n}$**) de (**a**) en función de los sentidos correctos.

1.5. Conversión del archivo de texto WordNet para español en el formato de base de datos.

- Para cada palabra y sus sinónimos se creó una entrada en una base de datos debidamente normalizada.

1.6. Preparación de la representación de las palabras usando las primitivas semánticas en formato de base de datos.

- Se usó un método y un programa de determinación de primitivas que para un diccionario dado permite detectar la lista de las primitivas.
- Se redefinieron las palabras del diccionario explicativo con la lista de las primitivas (excepto para las palabras que ya son primitivas).
- Para cada palabra y sus primitivas se creó una entrada en una base de datos debidamente normalizada.

2. Diseño del Método de desambiguación de sentidos de palabras.

2.1. Diseño del método de sustitución de las definiciones en profundidad y definición de la profundidad óptima.

- Se selecciona el documento a analizar (**d1**)
- Se obtiene el documento normalizado morfológicamente y sin homonimia morfológica (**d2**)
  - Se desarrollaron las heurísticas sintácticas. Ver Anexo 2. Algoritmo de desambiguación morfológica
  - Se realizó un análisis sintáctico parcial para las relaciones entre las palabras usando el software *parser sintáctico* que existe en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC. Es parcial por la posibilidad de fallar el acomplamiento sintáctico para todo el texto.
- Se obtuvo el documento sin palabras auxiliares (**d3**)
  - Se eliminaron las palabras auxiliares del documento (**d2**) como las preposiciones, verbos auxiliares, artículos y pronombres y se dejaron los verbos, adjetivos, sustantivos y adverbios.
- Se obtuvo el documento en *lemas* (**d4**)
  - Se normalizaron las palabras que conforman el documento (**d3**) y se sustituyeron todas las palabras que aparecen por sus ***lemas***. Ver Anexo

1. Algoritmo de normalización morfológica.

- Desambiguación de cada palabra del documento (**d3**) teniendo en cuenta un dominio local limitado a una ventana (**v**):
    - Se determinó el sentido correcto (**S**) de la palabra según su aporte en el diccionario explicativo desambiguado ANAYA en 1.3 y los diccionarios de sinónimos de Océano y WordNet desambiguados en 1.4

      Para todas las palabras del documento:

        - Seleccionar la palabra a analizar (**a**).
        - Seleccionar la definición de la palabra a analizar (**DA**)
        - Seleccionar los sinónimos de la palabra (**a**) , (**SA**)
        - Seleccionar las palabras de la ventana (**v**) ($\mathbf{v_{1....n}}$), limitado al párrafo donde se encuentra la palabra (**a**) exceptuando la propia palabra (**a**).
        - Seleccionar las definiciones de las palabras en **v**. ($\mathbf{V_{1..n}}$)
        - Seleccionar los sinónimos de las palabras en **v**. ($\mathbf{SV_{1..m}}$)
        - Seleccionar las palabras de cada una de las definiciones ($\mathbf{V_{1..n}}$). ($\mathbf{P_{V1..n}}$)
        - Escoger para la palabra (**a**) el sentido correcto:
            - Determinar el grado de semejanza entre **A** y $\mathbf{V_{1...n}}$ con un valor normalizado del peso entre **0** y **1** de las siguientes fuentes:
                - Determinar el aporte de la palabra (**a**) en cada definición $\mathbf{V_{1...n}}$ (**P1**)
                - Determinar el aporte de los sinónimos cercanos de la palabra (**a**) en cada definición $\mathbf{V_{1...n}}$ (**P2**)
                - Determinar el aporte de los sinónimos lejanos de la palabra (**a**) en cada definición $\mathbf{V_{1...n}}$ (**P3**)

- Determinar el aporte de los sinónimos cercanos de las palabras de la definición **V** (**$P_{V1...m}$**) en la definición **A**. (**P4**)
- Determinar el aporte de los sinónimos lejanos de las palabras de la definición **V** (**$P_{V1...m}$**) en la definición **A**. (**P5**)
- El Peso General del sentido en el nivel 1 es determinado por el factor ponderado de los pesos de las fuentes y su nivel de confianza en el contexto:

$$PG_i = 1 - \prod_{i=1}^{N} (1 - P_i)^{Cci}$$

Donde **N** es el número de fuentes, en este caso 5

P_i es el peso obtenido para cada fuente

Cc es la confianza en el contexto

$Cc_i = (Pd_i + Po_i + Pp_i)$

Donde Pd_i = Es la confianza respecto a la distancia

Po_i = Es la confianza respecto a la oración

Pp_i = Es la confianza respecto al párrafo

Cci = Es la confianza total

Determinados como sigue:

$$Pd_i = D^{-d^a}$$

$$Po_i = O^{-d^b}$$

Donde:

D, O y P: son los parámetros de anchura (*)
a, b, c: son los parámetros de aplanamiento (*)
d: es la distancia

(*) Estos parámetros se determinaron experimentalmente

$$Pp_i = P^{-d^C}$$

- **Determinar la profundidad recursiva en un nivel x** con un valor entre 0 y N, limitado por la relación entre los aportes recursivos que se consideren significativos.

  El Peso del sentido es determinado por la sumatoria ponderada de los pesos finales de cada profundidad, el que disminuye su confianza exponencialmente:

$$Peso = 1 - \sum_{i=1}^{N} (1 - Pf_i)^{Coi}$$

  Donde N es el número de profundidad

  Pf_i es el peso final obtenido para cada profundidad

  Co_i es la confianza

  La confianza de cada profundidad se normaliza exponencialmente de la siguiente manera:

$$\underset{i=1}{Co}^{N} = 1 - Ap^{-N}$$

  Donde se cree que la aportaciones disminuyen a la mitad en cada nivel, es decir, $A_P=2$.

  Para obtener una confianza deseable del 97%, se debe usar una profundidad de hasta 5.

  $Co=1- 2^{-5}=1- (1/2x1/2x1/2x1/2x1/2)=1-1/32=0.97$

  Con este factor de aportación $A_p=2$, y con una **profundidad de N=5** se puede obtener una

precisión del 97%.

Para cada nivel la confianza sería de:

N=1, Co = 0.5

N=2, Co = 0.25

N=3, Co = 0.12

N=4, Co = 0.06

N=5, Co = 0.03

- Determinar el aporte recursivo del grado de semejanza entre $V_{1...n}$ y las definiciones de cada palabra de $P_{V1...m}$ ($C_{1...m}$) con un valor normalizado entre 0 y 1 aplicando nuevamente el mismo algoritmo que se aplicó a la semejanza entre A y $V_{1...n,}$ deteniendo la profundidad recursiva en el **nivel 5**.

  Como se explicó anteriormente, ese aporte quedaría

$$\text{Peso} = 1 - \sum_{i=1}^{5} (1 - Pf_i)^{Coi}$$

  Donde 5 es el número de profundidad

  Pf_i es el peso final obtenido para cada profundidad

  Co_i es la confianza general al 97%

- Escoger como sentido correcto de la palabra (a) en el entorno local v, el que mayor peso obtenga.

- Obtener el documento (d5) al sustituir cada palabra del documento (d3) respectivamente, por la misma palabra con su sentido correcto.

2.2. Diseño del método de ponderación de los diferentes recursos léxicos del diccionario explicativo ANAYA y los diccionarios de sinónimos Océano y WordNet.

- Se probó el método con cada recurso por separado y se determinó cuál recurso influye con más calidad en el resultado.
- Se tomaron como referencia para el análisis de la calidad del proceso algunos documentos desambiguados manualmente.
- La relación obtenida es el factor de ponderación usado para cada recurso.

3. Preparación de la colección de documentos para hacer pruebas de funcionamiento del método.

3.1. Se determinaron los criterios de representatividad de los textos.

3.2. Se compiló la colección de documentos de Internet y de los corpus que existen en el Laboratorio de Lenguaje Natural del CIC-IPN.

3.3. Se realizó el marcado manual de algunas partes de los textos para los experimentos.

4. Análisis del funcionamiento del método y determinación el tamaño óptimo del contexto a usar para la desambiguación.

Al considerar una eficiencia del 97% se limita de forma dinámica la ventana del contexto según los siguientes criterios:

- Tomar una ventana **V** de **X** palabras donde la cantidad de palabras irá incrementándose por la izquierda y la derecha. La cantidad de palabras por el lado izquierdo se incrementa mientras el valor del error sea mayor a 1,5%; igualmente para el lado derecho, pero independiente uno del otro.
- El valor del error se comprobó teniendo en cuenta el resultado de las últimas 3 palabras analizadas de cada lado.
- El valor de los pesos según el contexto se calculó teniendo en cuenta que:
    - Las palabras muy cerca dentro de una distancia (d), **influyen mucho** sobre la palabra analizada (Pd)

- Las palabras que se encuentran en otra oración distinta a la palabra analizada influyen **en menor medida** que las palabras de la propia oración (Po)
- Las palabras que se encuentran en otro párrafo distinto a la palabra analizada influyen **en muchísima menor medida** que las palabras del propio párrafo. (Pp)
- La determinación de los pesos según el contexto se determinó exponencialmente como sigue:

$$Pd_i = D^{-d^a}$$

$$Po_i = O^{-d^b}$$

$$Pp_i = P^{-d^c}$$

Donde:

D, O y P: son los parámetros de anchura (*)
a, b, c: son los parámetros de aplanamiento (*)
d: es la distancia

(*) Estos parámetros se determinaron experimentalmente

Donde

Pd_i es la confianza respecto a la distancia;

Po_i es la confianza respecto a la oración;

Pp_i es la confianza respecto al párrafo;

PG_i es el peso general según las fuentes.

5. **Realización de las pruebas de eficiencia del método en una colección de documentos en español.**

5.1. Se escogieron dos métodos como "línea base" (método original de Lesk, método modificado de Lesk).

5.2. Se realizaron las pruebas.

5.3. Se compararon los resultados.

2.5 Aplicación práctica del método

El nuevo método de recuperación de información con resolución de la ambigüedad de los sentidos de las palabras puede usarse en:

- Buscadores de información en Internet.
- Dentro de un sistema de recuperación de información en bibliotecas digitales tradicionales.
- Para generar bibliotecas digitales indexadas según los sentidos de palabras.
- Para buscar información relevante dentro de una colección de documentos locales.
- Para refinar búsquedas realizadas en los buscadores que existen actualmente en Internet.

2.6 Límites y limitaciones

El nuevo Método de desambiguación de sentidos de palabras para el español, sólo está preparado para información textual, por lo que no se usaría para desambiguar información contenida en documentos de multimedia como voz y vídeo.

El método está preparado para documentos textuales escritos solamente en español, por lo que no se puede usar en documentos redactados en otros idiomas, aunque no se descarta la posibilidad de su funcionamiento para otros idiomas siempre y cuando se dispongan de las fuentes necesarias y se realicen los análisis correspondientes.

3

Resultados

En este capítulo se presentan los resultados y la comparación del método con otras soluciones así cómo los aportes realizados.

Se mencionan los méritos y las publicaciones durante el desarrollo del doctorado.

Capítulo 3. Resultados

3.1 Programación del método

La programación del método se realizó en Borlad C. Este programa permite realizar las pruebas de desambiguación de los sentidos de las palabras por tres métodos, el método original de Lesk, el método de Lesk modificado y el método propuesto en este trabajo.

3.1.1 Programación del Método de Lesk Original

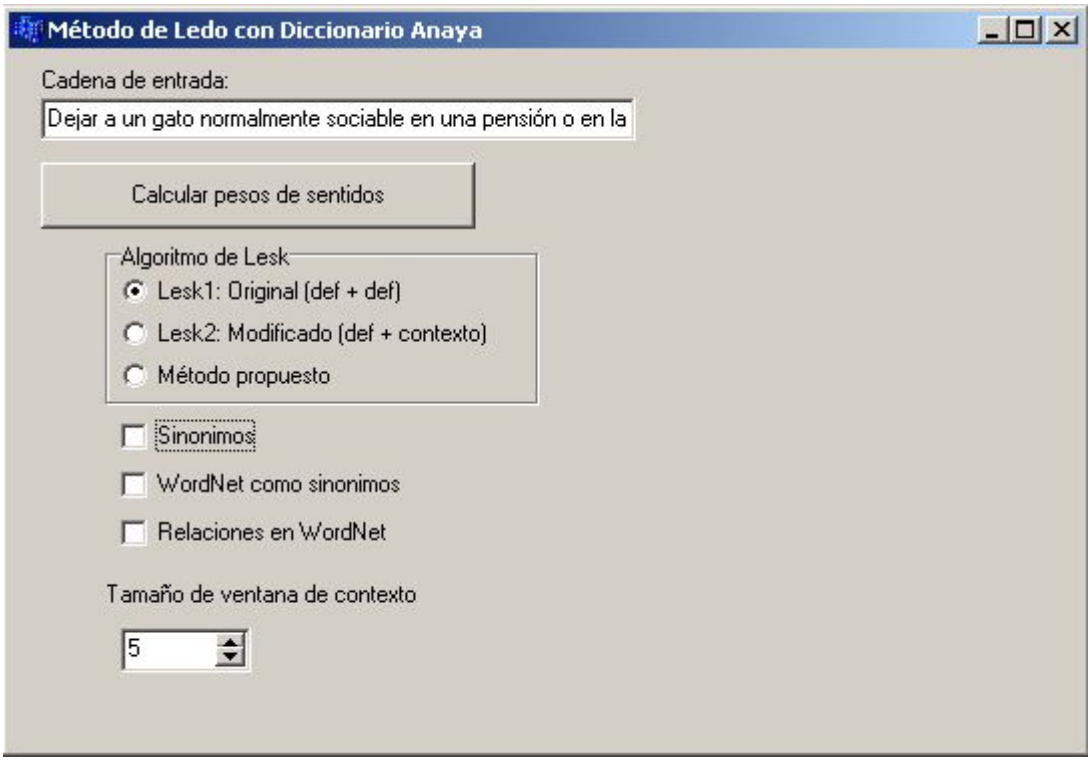


Figura 7. Pantalla del programa para el método de Lesk original

3.1.2 Programación para el Método de Lesk Modificado

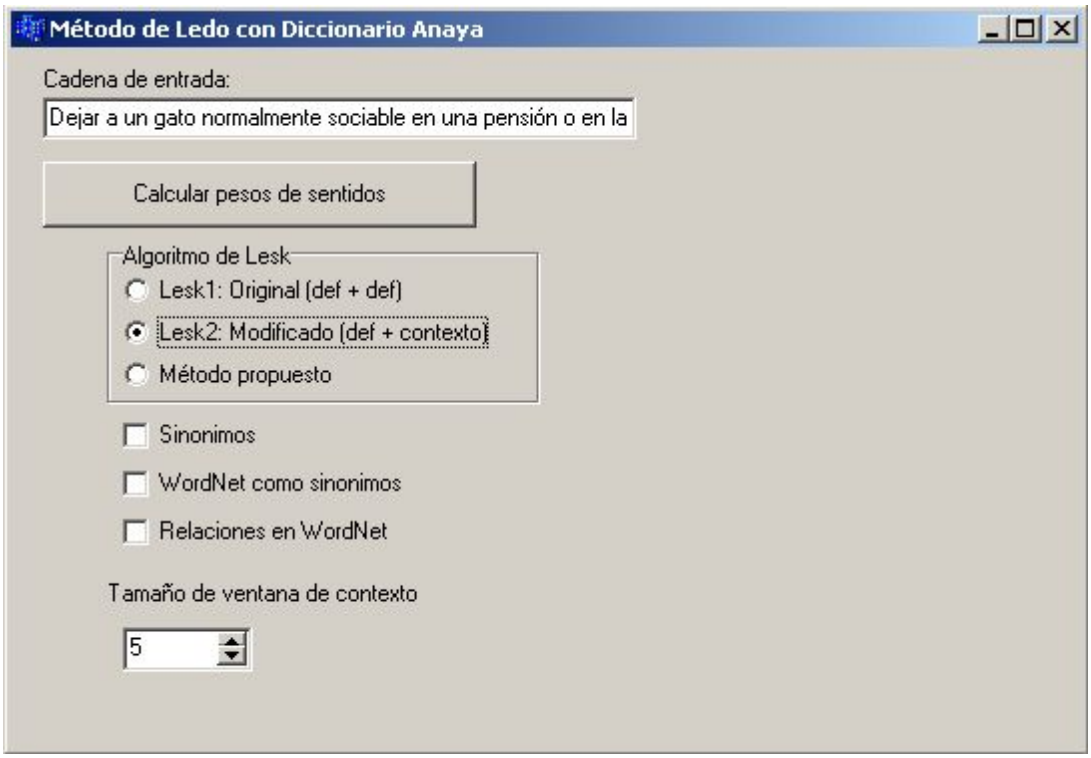


Figura 8. Pantalla del programa para el método de Lesk modificado

3.1.3 Programación para el Método Propuesto

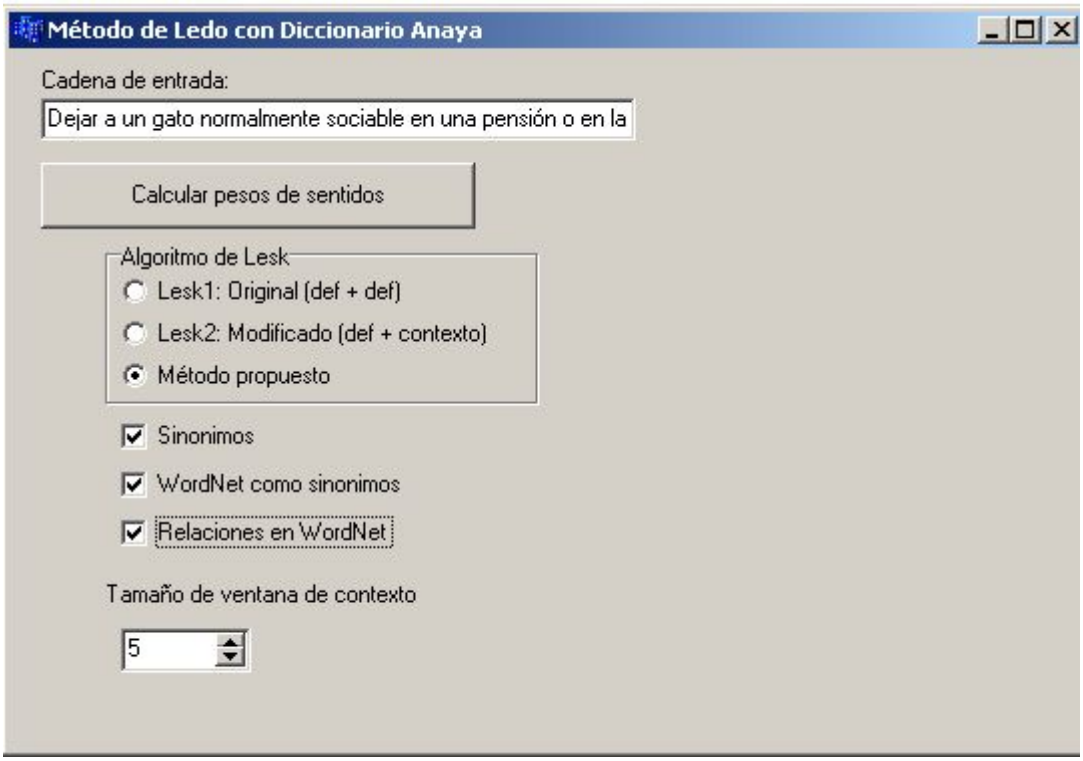


Figura 9. Pantalla del programa para el método propuesto

3.2 Ejemplo del procesamiento con el programa

3.2.1 Ejemplo para el Método de Lesk Original

Dejar a un gato normalmente sociable en una pensión o en la clínica.

words, window sz: 6 5

Word 0: dejar # sences 14
gato 7 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0

pensión 4 Sence 0: []0 Sence 1: [E:no]0,0240437 Sence 2: []0 Sence 3: []0 Sence 4: [E:no]0,0240437 Sence 5: []0 Sence 6: [E:no]0,0240437 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
Word 1: gato # sences 7
dejar 14 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: [E:persona]0,31427 Sence 5: [E:persona]0,31427 Sence 6: []0
pensión 4 Sence 0: [E:tener]0,0279363 Sence 1: []0 Sence 2: [E:tener]0,0279363 Sence 3: []0 Sence 4: [E:persona]0,0279363 Sence 5: [E:persona]0,0279363 Sence 6: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
Word 2: normalmente # sences 0
dejar 14
gato 7
sociable 1
pensión 4
clínica 5
Word 3: sociable # sences 1
dejar 14 Sence 0: []0
gato 7 Sence 0: [E:persona]0,0214275
normalmente 0 Sence 0: []0
pensión 4 Sence 0: [E:persona]0,0318069
clínica 5 Sence 0: []0
Word 4: pensión # sences 4
dejar 14 Sence 0: []0 Sence 1: [E:no]0,0324068 Sence 2: []0 Sence 3: []0
gato 7 Sence 0: [E:tener]0,0196824 Sence 1: [E:persona]0,0196824 Sence 2: []0 Sence 3: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
sociable 1 Sence 0: []0 Sence 1: [E:persona]0,140859 Sence 2: []0 Sence 3: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
Word 5: clínica # sences 5
dejar 14 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0
gato 7 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0

normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0
pensión 4 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0
Tracing done.

Word 0: dejar # sences 14
Sence 0: 0,000000 Soltar una cosa; apartarse de algo o de alguien
Sence 1: 0,000000 No impedir
Sence 2: 0,000000 Desamparar
Sence 3: 0,000000 Encargar
Sence 4: 0,000000 No proseguir lo comenzado
Sence 5: 0,000000 Prestar
Sence 6: 0,000000 No perturbar ni molestar
Sence 7: 0,000000 Legar
Sence 8: 0,000000 Abandonarse
Sence 9: 0,000000 Abatirse, caer de ánimo
Sence 10: 0,000000 Omitir
Sence 11: 0,000000 Producir ganancia
Sence 12: 0,000000 Nombrar
Sence 13: 0,000000 Ausentarse
Word 1: gato # sences 7
Sence 0: 0,000000 Nombre común con el que se conoce cierto mamífero carnívoro doméstico, de la familia de los félidos, que se tiene como animal de compañía (dependiendo de las clasificaciones, <I>Felis sylvestris</I> o <I>Felis catus</I>).
Sence 1: 0,000000 Utensilio para elevar a poca altura grandes pesos, especialmente el utilizado para levantar los automóviles.
Sence 2: 0,000000 Nombre común que reciben los mamíferos carnívoros de la familia de los félidos, especialmente los del género <I>Felis,</I> que se caracterizan por tener orejas cortas, cola larga, uñas retráctiles y visión muy aguda.
Sence 3: 0,000000 Ladrón.
Sence 4: 0,000000 Persona astuta.
Sence 5: 0,000000 Persona natural de Madrid.
Sence 6: 0,000000 Mercado al aire libre.
Word 2: normalmente # sences 0
Word 3: sociable # sences 1
Sence 0: 0,000000 Dícese de la persona de trato fácil.
Word 4: pensión # sences 4
Sence 0: 0,000000 Casa particular que tiene un número normalmente reducido y permanente de huéspedes, y cuyo precio es inferior al del hotel.
Sence 1: 0,000000 Asignación que disfruta una persona por un trabajo que ya no realiza en la actualidad
Sence 2: 0,000000 Renta anual que se impone sobre una finca.

Sence 3: 0,000000 Cantidad que se paga mensualmente, o al año, por algún servicio propio o ajeno.
Word 5: clínica # sences 5
Sence 0: 0,000000 Hospital privado.
Sence 1: 0,000000 Parte práctica de la enseñanza médica.
Sence 2: 0,000000 Departamento hospitalario donde se imparte esta enseñanza.
Sence 3: 0,000000 Relativo a la clínica.
Sence 4: 0,000000 Dícese del especialista en la práctica clínica.
Presentation done.

3.2.2 Ejemplo para el Método de Lesk Modificado

Dejar a un gato normalmente sociable en una pensión o en la clínica.

words, window sz: 6 5

Word 0: dejar # sences 14
gato 7 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
pensión 4 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0 Sence 7: []0 Sence 8: []0 Sence 9: []0 Sence 10: []0 Sence 11: []0 Sence 12: []0 Sence 13: []0
Word 1: gato # sences 7
dejar 14 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
pensión 4 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0 Sence 4: []0 Sence 5: []0 Sence 6: []0
Word 2: normalmente # sences 0
dejar 14
gato 7

sociable 1
pensión 4
clínica 5
Word 3: sociable # sences 1
dejar 14 Sence 0: []0
gato 7 Sence 0: []0
normalmente 0 Sence 0: []0
pensión 4 Sence 0: []0
clínica 5 Sence 0: []0
Word 4: pensión # sences 4
dejar 14 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
gato 7 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0
Word 5: clínica # sences 5
dejar 14 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: [
E:clínica]0,0921285 Sence 4: [E:clínica]0,0921285
gato 7 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: [
E:clínica]0,124226 Sence 4: [E:clínica]0,124226
normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: [
E:clínica]0,144338 Sence 4: [E:clínica]0,144338
sociable 1 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: [
E:clínica]0,157135 Sence 4: [E:clínica]0,157135
pensión 4 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: [
E:clínica]0,164336 Sence 4: [E:clínica]0,164336
Tracing done.

Word 0: dejar # sences 14
Sence 0: 0,000000 Soltar una cosa; apartarse de algo o de alguien
Sence 1: 0,000000 No impedir
Sence 2: 0,000000 Desamparar
Sence 3: 0,000000 Encargar
Sence 4: 0,000000 No proseguir lo comenzado
Sence 5: 0,000000 Prestar
Sence 6: 0,000000 No perturbar ni molestar
Sence 7: 0,000000 Legar
Sence 8: 0,000000 Abandonarse
Sence 9: 0,000000 Abatirse, caer de ánimo
Sence 10: 0,000000 Omitir
Sence 11: 0,000000 Producir ganancia
Sence 12: 0,000000 Nombrar
Sence 13: 0,000000 Ausentarse
Word 1: gato # sences 7

Sence 0: 0,000000 Nombre común con el que se conoce cierto mamífero carnívoro doméstico, de la familia de los félidos, que se tiene como animal de compañía (dependiendo de las clasificaciones, <I>Felis sylvestris</I> o <I>Felis catus</I>).
Sence 1: 0,000000 Utensilio para elevar a poca altura grandes pesos, especialmente el utilizado para levantar los automóviles.
Sence 2: 0,000000 Nombre común que reciben los mamíferos carnívoros de la familia de los félidos, especialmente los del género <I>Felis,</I> que se caracterizan por tener orejas cortas, cola larga, uñas retráctiles y visión muy aguda.
Sence 3: 0,000000 Ladrón.
Sence 4: 0,000000 Persona astuta.
Sence 5: 0,000000 Persona natural de Madrid.
Sence 6: 0,000000 Mercado al aire libre.
Word 2: normalmente # sences 0
Word 3: sociable # sences 1
Sence 0: 0,000000 Dícese de la persona de trato fácil.
Word 4: pensión # sences 4
Sence 0: 0,000000 Casa particular que tiene un número normalmente reducido y permanente de huéspedes, y cuyo precio es inferior al del hotel.
Sence 1: 0,000000 Asignación que disfruta una persona por un trabajo que ya no realiza en la actualidad
Sence 2: 0,000000 Renta anual que se impone sobre una finca.
Sence 3: 0,000000 Cantidad que se paga mensualmente, o al año, por algún servicio propio o ajeno.
Word 5: clínica # sences 5
Sence 0: 0,000000 Hospital privado.
Sence 1: 0,000000 Parte práctica de la enseñanza médica.
Sence 2: 0,000000 Departamento hospitalario donde se imparte esta enseñanza.
Sence 3: 0,164336 Relativo a la clínica.
Sence 4: 0,164336 Dícese del especialista en la práctica clínica.
Presentation done.

3.2.3 Ejemplo para el Método Propuesto

Dejar a un gato normalmente sociable en una pensión o en la clínica.

words, window sz: 6 5

Word 0: dejar # sences 14
gato 7 Sence 0: []0,231564 Sence 1: []0,164336 Sence 2: []0,164336 Sence 3: []0,164336 Sence 4: []0 Sence 5: []0,164336 Sence 6: []0 Sence 7: []0,164336 Sence 8: []0 Sence 9: []0,164336 Sence 10: []0,186745 Sence 11: []0,164336 Sence 12: []0,0448188 Sence 13: []0
normalmente 0 Sence 0: []0,157135 Sence 1: []0,157135 Sence 2: []0,157135 Sence 3: []0,157135 Sence 4: []0 Sence 5: []0,157135 Sence 6:

[]0 Sence 7: []0,157135 Sence 8: []0 Sence 9: []0,157135 Sence 10: []0,157135 Sence 11: []0,157135 Sence 12: []0 Sence 13: []0

sociable 1 Sence 0: []0,360844 Sence 1: []0,144338 Sence 2: []0,144338 Sence 3: []0,144338 Sence 4: []0 Sence 5: []0,144338 Sence 6: []0 Sence 7: []0,144338 Sence 8: []0 Sence 9: []0,144338 Sence 10: []0,144338 Sence 11: []0,144338 Sence 12: []0 Sence 13: []0

pensión 4 Sence 0: []0,172313 Sence 1: [E:no]0,14827 Sence 2: []0,124226 Sence 3: []0,124226 Sence 4: [E:no]0,0480875 Sence 5: []0,124226 Sence 6: [E:no]0,0480875 Sence 7: []0,124226 Sence 8: []0 Sence 9: []0,124226 Sence 10: []0,124226 Sence 11: []0,172313 Sence 12: []0,0480875 Sence 13: []0

clínica 5 Sence 0: []0,138193 Sence 1: []0,0921285 Sence 2: []0,0921285 Sence 3: []0,0921285 Sence 4: []0 Sence 5: []0,0921285 Sence 6: []0 Sence 7: []0,0921285 Sence 8: []0 Sence 9: []0,0921285 Sence 10: []0,0921285 Sence 11: []0,0921285 Sence 12: []0 Sence 13: []0

Word 1: gato # sences 7

dejar 14 Sence 0: []0,0857403 Sence 1: []0,12861 Sence 2: []0,0857403 Sence 3: []0,164336 Sence 4: []0,164336 Sence 5: []0 Sence 6: []0,164336

normalmente 0 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0,164336 Sence 4: []0,164336 Sence 5: []0 Sence 6: []0,164336

sociable 1 Sence 0: []0 Sence 1: []0,104757 Sence 2: []0,0471405 Sence 3: []0,157135 Sence 4: [E:persona]0,471405 Sence 5: [E:persona]0,31427 Sence 6: []0,471405

pensión 4 Sence 0: [E:tener]0,167618 Sence 1: []0,167618 Sence 2: [E:tener]0,251427 Sence 3: []0,144338 Sence 4: [E:persona]0,172274 Sence 5: [E:persona]0,111745 Sence 6: []0,172274

clínica 5 Sence 0: []0 Sence 1: []0 Sence 2: []0 Sence 3: []0,124226 Sence 4: []0,186339 Sence 5: []0,062113 Sence 6: []0,186339

Word 2: normalmente # sences 0

dejar 14

gato 7

sociable 1

pensión 4

clínica 5

Word 3: sociable # sences 1

dejar 14 Sence 0: []0,0376533

gato 7 Sence 0: [E:persona]0,042855

normalmente 0 Sence 0: []0

pensión 4 Sence 0: [E:persona]0,0318069

clínica 5 Sence 0: []0,0785674

Word 4: pensión # sences 4

dejar 14 Sence 0: []0,0897419 Sence 1: [E:no]0,129627 Sence 2: []0,0648136 Sence 3: []0,0648136

gato 7 Sence 0: [E:tener]0,283124 Sence 1: [E:persona]0,0590472 Sence 2: []0,0393648 Sence 3: []0,098412

normalmente 0 Sence 0: []0,0725238 Sence 1: []0 Sence 2: []0 Sence 3: []0
sociable 1 Sence 0: []0,151694 Sence 1: [E:persona]0,140859 Sence 2: []0 Sence 3: []0
clínica 5 Sence 0: []0,151694 Sence 1: []0,0821678 Sence 2: []0 Sence 3: []0,0821678
Word 5: clínica # sences 5
dejar 14 Sence 0: []0,0921285 Sence 1: []0,0240335 Sence 2: []0 Sence 3: [E:clínica]0,0921285 Sence 4: [E:clínica]0,0921285
gato 7 Sence 0: []0,124226 Sence 1: []0,0338798 Sence 2: []0 Sence 3: [E:clínica]0,124226 Sence 4: [E:clínica]0,124226
normalmente 0 Sence 0: []0,144338 Sence 1: []0 Sence 2: []0 Sence 3: [E:clínica]0,144338 Sence 4: [E:clínica]0,144338
sociable 1 Sence 0: []0,157135 Sence 1: []0,235702 Sence 2: []0 Sence 3: [E:clínica]0,157135 Sence 4: [E:clínica]0,157135
pensión 4 Sence 0: []0,196142 Sence 1: []0,0954206 Sence 2: []0 Sence 3: [E:clínica]0,164336 Sence 4: [E:clínica]0,164336
Tracing done.

Word 0: dejar # sences 14
Sence 0: 0,138193 Soltar una cosa; apartarse de algo o de alguien
Sence 1: 0,092128 No impedir
Sence 2: 0,092128 Desamparar
Sence 3: 0,092128 Encargar
Sence 4: 0,000000 No proseguir lo comenzado
Sence 5: 0,092128 Prestar
Sence 6: 0,000000 No perturbar ni molestar
Sence 7: 0,092128 Legar
Sence 8: 0,000000 Abandonarse
Sence 9: 0,092128 Abatirse, caer de ánimo
Sence 10: 0,092128 Omitir
Sence 11: 0,092128 Producir ganancia
Sence 12: 0,000000 Nombrar
Sence 13: 0,000000 Ausentarse
Word 1: gato # sences 7
Sence 0: 0,000000 Nombre común con el que se conoce cierto mamífero carnívoro doméstico, de la familia de los félidos, que se tiene como animal de compañía (dependiendo de las clasificaciones, <I>Felis sylvestris</I> o <I>Felis catus</I>).
Sence 1: 0,000000 Utensilio para elevar a poca altura grandes pesos, especialmente el utilizado para levantar los automóviles.
Sence 2: 0,000000 Nombre común que reciben los mamíferos carnívoros de la familia de los félidos, especialmente los del género <I>Felis,</I> que se caracterizan por tener orejas cortas, cola larga, uñas retráctiles y visión muy aguda.
Sence 3: 0,124226 Ladrón.
Sence 4: 0,186339 Persona astuta.
Sence 5: 0,062113 Persona natural de Madrid.

Sence 6: 0,186339 Mercado al aire libre.
Word 2: normalmente # sences 0
Word 3: sociable # sences 1
Sence 0: 0,078567 Dícese de la persona de trato fácil.
Word 4: pensión # sences 4
Sence 0: 0,151694 Casa particular que tiene un número normalmente reducido y permanente de huéspedes, y cuyo precio es inferior al del hotel.
Sence 1: 0,082168 Asignación que disfruta una persona por un trabajo que ya no realiza en la actualidad
Sence 2: 0,000000 Renta anual que se impone sobre una finca.
Sence 3: 0,082168 Cantidad que se paga mensualmente, o al año, por algún servicio propio o ajeno.
Word 5: clínica # sences 5
Sence 0: 0,196142 Hospital privado.
Sence 1: 0,095421 Parte práctica de la enseñanza médica.
Sence 2: 0,000000 Departamento hospitalario donde se imparte esta enseñanza.
Sence 3: 0,164336 Relativo a la clínica.
Sence 4: 0,164336 Dícese del especialista en la práctica clínica.
Presentation done.

3.3 Experimentos realizados

Se analizaron **4,287** significados pertenecientes a 872 palabras desambiguadas usadas en 80 contextos con 1,293 palabras, usando los tres métodos (Lesk original, Lesk modificado, Método propuesto)

No.	Contexto	Palabras desambiguadas	Significados analizados
1.	Dejar a un gato normalmente sociable en una pensión o en la clínica.	6	31
2.	Detienen a militar implicado en el asesinato de policías	6	14
3.	No se pierda las razones para suscribirse al mejor diario digital	9	50
4.	Participa en nuestro concurso de fotografía digital	5	12
5.	El Ejecutivo espera que en el pacto participen todos los partidos democráticos	9	33
6.	No guarde su contraseña si accede desde un lugar público	9	34
7.	La visita de la vicepresidenta primera del Gobierno a la ciudad fue un éxito	8	59
8.	El servicio más económico para enviar dinero a sus familiares	8	21
9.	Más de 5000 soldados desertaron de las fuerzas	9	37

	armadas estadounidenses desde que comenzó la invasión		
10.	La vacuna terapéutica contra el cáncer de pulmón pasará en el 2005 a fase de ensayo clínico	11	66
11.	El programa oficial distribuido por la Cancillería indicaba el inicio del evento	10	32
12.	Todas las maniobras serán evaluadas por experimentados especialistas que actuarán como árbitros	8	41
13.	Algunos de los inventos más importantes para la Humanidad han surgido de ideas muy sencillas	10	45
14.	Las pinturas y esculturas de las tumbas egipcias son las primeras representaciones de la humanidad	8	47
15.	Tu perro no me deja en paz ni un rato en plena reunión familiar	11	40
16.	La Unión Europea tiene interés en armonizar las leyes de todos los países	8	36
17.	El nadador estadounidense gana dos pruebas más y suma ya cinco medallas en Atenas	12	57
18.	La revista era sumamente interesante e instructiva	5	32
19.	La mayoría de las veces que uno acude a una tienda fotográfica el vendedor nos recomienda un rollo	9	37
20.	Quienes compran seguro de vida generalmente lo consideran una inversión	7	26
21.	Se reforzará la presencia policiaca en vías primarias	6	18
22.	El operativo será aplicado en los centros comerciales con mayor afluencia	8	50
23.	El accidente ha tenido lugar sobre el pueblo	7	47
24.	Defensores de los derechos de los ancianos argumentan que las personas de la tercera edad se sienten cada vez más acorraladas	12	62
25.	Las personas de la tercera edad notan un declive en los ingresos que reciben	7	29
26.	De noche, cientos de miles de luces iluminan casas, edificios y árboles, a la vez que inmensos Santa Claus saludan desde las esquinas	13	67
27.	Este año, las piezas favoritas para la decoración exterior son las figuras holográficas en forma de palmeras	10	84
28.	Pensó que había pasado sobre unos tablones que había en la entrada del garaje, frenó, dio adelante y volvió a retroceder	13	172
29.	La guerrilla de las Fuerzas Armadas Revolucionarias son de extrema izquierda	7	65
30.	Participaron en el certamen físicos recién graduados de diversas partes del país	10	38

31.	Uno de los principales aportes del cristianismo a la cultura fue la introducción de la historia como elemento que dinamiza a la religión	11	75
32.	Propone gobierno capitalino a nuevo Secretario de Seguridad Pública	6	36
33.	Descubren documentos que dan nuevos datos sobre la conquista	6	58
34.	La Orquesta Filarmónica y el Coro Juvenil de la Ciudad de México interpretarán villancicos de diferentes partes del mundo	13	57
35.	Se premiará un producto de investigación que haga una aportación de calidad a las Ciencias Sociales	9	50
36.	El Congreso tuvo como objetivo principal el dar a conocer los nuevos avances científicos y tecnológicos en el área de Cómputo	13	88
37.	El taller de Sistemas de Información tuvo como tema central el Control de la Calidad de Software	11	42
38.	Se mostró un prototipo de una Urna Electrónica	5	10
39.	Al concluir la presentación se propusieron 2 proyectos más para trabajar en colaboración	9	23
40.	A inicios de diciembre se llevó a cabo el tercer seminario relativo a la Organización Mundial de Comercio	12	48
41.	Agricultores y ganaderos piden al Ministerio ayuda por la sequía como las concedidas por la ola de frío	10	38
42.	La Policía desarticula en Gran Canaria una red de tráfico de menores africanos destinados a la prostitución	10	31
43.	Dos explosiones de origen desconocido se produjeron hoy en el exterior del consulado británico en Nueva York, pero según fuentes policiales no hubo que lamentar víctimas	22	87
44.	Hoy parece claro que la meta es la competitividad, pues sin ella el horizonte se llena de incertidumbre	13	67
45.	Un joven ingeniero francés estudiante en China me decía que su problema no era el chino sino las matemáticas	15	67
46.	Ayer defendimos nuestra soberanía con las armas, hoy nos toca defender nuestra democracia para consolidarla pero, sobre todo, para mejorarla	16	71
47.	Con la intervención francesa, expuso, se corría el riesgo de perder por entero la soberanía nacional	11	65
48.	Democracia implica discusiones y pactos, implica también diferencias y acuerdos, precisó el secretario de Gobernación	12	34
49.	En la conferencia de prensa mañanera, anunció que el	13	39

	lunes realizará una ceremonia simbólica para concluir la resistencia civil pacífica		
50.	Bill Gates, presidente de Microsoft, asegura que la siguiente generación de consolas de juego de Microsoft, Xbox, dará al más grande productor de software la oportunidad de quitarle el liderazgo a Sony y tomar el negocio de los juegos	26	123
51.	La relación valor-precio del seguro ante inundaciones, terremoto, incendio y explosión, entre otros siniestros, es muy alta para el asegurado	17	101
52.	Impartir educación a nivel postgrado en las áreas de Ciencias de la Computación e Ingeniería de Cómputo	10	26
53.	Para el Centro de Investigación en Computación, la vinculación con los diferentes sectores de la sociedad, representa el eje sobre el que se apoyarán las acciones educativas y de investigación	18	70
54.	Tony Blair se encamina a obtener un tercer mandato en las elecciones de este jueves en Gran Bretaña. Los sondeos le dan a Blair 14 puntos de ventaja sobre su rival más próximo	22	82
55.	Se cree que el terrorista, de origen libio, es uno de los pocos que conoce el paradero de Bin Laden	13	39
56.	Tony Blair ganó ayer un histórico tercer período como primer ministro de Gran Bretaña, hecho sin precedentes para el Partido Laborista	16	74
57.	Hubo un momento especialmente interesante en la conferencia de prensa conjunta que ofrecieron los presidentes de Estados Unidos, México y Canadá	13	72
58.	La toma de posesión del presidente socialista uruguayo Tabaré Vásquez ha sido interpretada, casi unánimemente, como la última evidencia de un giro hacia la izquierda en toda América Latina	21	102
59.	Algunas de las víctimas fueron identificadas por sus familiares, quienes dijeron que eran campesinos que habían desaparecido recientemente	13	91
60.	Las autoridades mantuvieron a los periodistas a distancia. Pero un fotógrafo de Associated Press vio a militares norteamericanos.	11	41
61.	En la celebración de ayer estuvieron presentes personalidades del mundo del espectáculo como el productor musical Emilio Estefan.	14	55
62.	El objetivo es integrar un frente católico común para lograr incidir en una reforma migratoria integral.	12	70
63.	Los científicos ya tienen certeza de que existió agua en ese planeta, la han hallado congelada en ríos y en los polos	13	75

64.	El volcán Láscar, en el norte del país, entró en violenta erupción y dejó una nube de cenizas	13	55
65.	Meteorólogos pronosticaron el viernes tormentas eléctricas con fuertes vientos y granizo	9	26
66.	España ha sufrido su sequía más grave desde que empezó a llevar estadísticas, con un clima seco y frío de casi medio año	17	111
67.	Durante el invierno, España tuvo un promedio de 70 milímetros de lluvia, cuando la precipitación normal es de 200 milímetros	13	46
68.	La batalla de México en contra de los violentos cárteles de la droga podría verse afectada por la salida del procurador general	13	67
69.	Los analistas dicen que la política por el momento parece haber desplazado a la guerra contra las drogas en las prioridades	12	66
70.	En los cursos efectuados en la Ciudad de México los participantes conocieron la forma en la cual se administran las redes	11	50
71.	Aquí hay gato encerrado	4	28
72.	El gato que está en nuestro cielo, nunca se olvida que fuiste mía	10	68
73.	Pásame el gato que el auto se averió	7	46
74.	Los gatos se enferman por situaciones de estrés, y se arrancan sus propios pelos	10	37
75.	Cuando el gato percibe una amenaza se tensa e hincha la cola	10	42
76.	Los gatos de cremallera son adecuados para la elevación de pesos pequeños	7	46
77.	El auto negro que va hacia Madrid va conducido por un gato	9	63
78.	En las afueras de Lima, el gato está lleno de quioscos y tenderetes donde comerciantes venden de todo	12	65
79.	Eres un gato para los negocios y no se puede confiar en ti	8	44
80.	Al robar el banco, la policía atrapo al gato y lo metieron en preso	10	38
Totales	**80 contextos con 1,293 palabras donde se desambiguaron 872 palabras con 4,287 sentidos de palabras**	**872**	**4,287**

3.3.1 Ejemplo de mediciones del contexto 1.

3.3.1.1 Contexto 1. Método de Lesk original.

Contexto "*Dejar a un gato normalmente sociable en una pensión o en la clínica*"

Contexto
Dejar a un gato normalmente sociable en una pensión o en la clínica.

# words	6	window sz	5

		Sences: 14													
Word 0	dejar	0	1	2	3	4	5	6	7	8	9	10	11	12	13
gato	7	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
pensión	4	0.0000000	0.0240437	0.0000000	0.0000000	0.0240437	0.0000000	0.0240437	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0240437	0.0000000	0.0000000	0.0240437	0.0000000	0.0240437	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

		Sences: 7						
Word 1	gato	0	1	2	3	4	5	6
dejar	14	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000	0.3142700	0.3142700	0.0000000
pensión	4	0.0279363	0.0000000	0.0279363	0.0000000	0.0279363	0.0279363	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0279363	0.0000000	0.0279363	0.0000000	0.3422063	0.3422063	0.0000000

		Sences: 0
Word 2	normalmente	
dejar	14	
gato	7	
sociable	1	
pensión	4	
clínica	5	
Totales		

		Sences: 4			
Word 4	pensión	0	1	2	3
dejar	14	0.0000000	0.0324068	0.0000000	0.0000000
gato	7	0.0196824	0.0196824	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.1408590	0.0000000	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0196824	0.1929482	0.0000000	0.0000000

		Sences: 1
Word 3	sociable	0
dejar	14	0.0000000
gato	7	0.0214275
normalmente	0	0.0000000
pensión	4	0.0318069
clínica	5	0.0000000
Totales		0.0532344

		Sences: 5				
Word 5	clínica	0	1	2	3	4
dejar	14	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
gato	7	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
pensión	4	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

3.3.1.2 Contexto 1. Método de Lesk Modificado.

Contexto "*Dejar a un gato normalmente sociable en una pensión o en la clínica*"

Contexto
Dejar a un gato normalmente sociable en una pensión o en la clínica.

# words	6	window sz	5

Word 0	dejar	Sences: 14													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
gato	7	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
pensión	4	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

Word 1	gato	Sences: 7						
		0	1	2	3	4	5	6
dejar	14	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
pensión	4	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

Word 2	normalmente	Sences: 0
dejar	14	
gato	7	
sociable	1	
pensión	4	
clínica	5	
Totales		

Word 4	pensión	Sences: 4			
		0	1	2	3
dejar	14	0.0000000	0.0000000	0.0000000	0.0000000
gato	7	0.0000000	0.0000000	0.0000000	0.0000000
normalmente	0	0.0000000	0.0000000	0.0000000	0.0000000
sociable	1	0.0000000	0.0000000	0.0000000	0.0000000
clínica	5	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000

Word 3	sociable	Sences: 1
		0
dejar	14	0.0000000
gato	7	0.0000000
normalmente	0	0.0000000
pensión	4	0.0000000
clínica	5	0.0000000
Totales		0.0000000

Word 5	clínica	Sences: 5				
		0	1	2	3	4
dejar	14	0.0000000	0.0000000	0.0000000	0.0921285	0.0921285
gato	7	0.0000000	0.0000000	0.0000000	0.1242260	0.1242260
normalmente	0	0.0000000	0.0000000	0.0000000	0.1443380	0.1443380
sociable	1	0.0000000	0.0000000	0.0000000	0.1571350	0.1571350
pensión	4	0.0000000	0.0000000	0.0000000	0.1643360	0.1643360
Totales		0.0000000	0.0000000	0.0000000	0.6821635	0.6821635

3.3.1.3 Contexto 1. Método propuesto.

Contexto "*Dejar a un gato normalmente sociable en una pensión o en la clínica*"

Contexto
Dejar a un gato normalmente sociable en una pensión o en la clínica.

# words	6	window sz	5

Word 0	dejar	Sences: 14													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
gato	7	0.2315640	0.1643360	0.1643360	0.1643360	0.0000000	0.1643360	0.0000000	0.1643360	0.0000000	0.1643360	0.1867450	0.1643360	0.0448188	0.0000000
normalmente	0	0.1571350	0.1571350	0.1571350	0.1571350	0.0000000	0.1571350	0.0000000	0.1571350	0.0000000	0.1571350	0.1571350	0.1571350	0.0000000	0.0000000
sociable	1	0.3608440	0.1443380	0.1443380	0.1443380	0.0000000	0.1443380	0.0000000	0.1443380	0.0000000	0.1443380	0.1443380	0.1443380	0.0000000	0.0000000
pensión	4	0.1723130	0.1482700	0.1242260	0.1242260	0.0480875	0.1242260	0.0480875	0.1242260	0.1242260	0.1242260	0.1242260	0.1723130	0.0480875	0.0000000
clínica	5	0.1381930	0.0921285	0.0921285	0.0921285	0.0000000	0.0921285	0.0000000	0.0921285	0.0000000	0.0921285	0.0921285	0.0921285	0.0000000	0.0000000
Totales		1.0600490	0.7062075	0.6821635	0.6821635	0.0480875	0.6821635	0.0480875	0.6821635	0.1242260	0.6821635	0.7045725	0.7302505	0.0929063	0.0000000

Word 1	gato	Sences: 7						
		0	1	2	3	4	5	6
dejar	14	0.0857403	0.1286100	0.0857403	0.1643360	0.1643360	0.0000000	0.1643360
normalmente	0	0.0000000	0.0000000	0.0000000	0.1643360	0.1643360	0.0000000	0.1643360
sociable	1	0.0000000	0.1047570	0.0471405	0.1571350	0.4714050	0.3142700	0.4714050
pensión	4	0.1676180	0.1676180	0.2514270	0.1443380	0.1722740	0.1117450	0.1722740
clínica	5	0.0000000	0.0000000	0.0000000	0.1242260	0.1863390	0.0621130	0.1863390
Totales		0.2533583	0.4009850	0.3843078	0.7543710	1.1586900	0.4881280	1.1586900

Word 2	normalmente	Sences: 0
dejar	14	
gato	7	
sociable	1	
pensión	4	
clínica	5	
Totales		

Word 4	pensión	Sences: 4			
		0	1	2	3
dejar	14	0.0897419	0.1296270	0.0648136	0.0648136
gato	7	0.2831240	0.0590472	0.0393648	0.0984120
normalmente	0	0.0725238	0.0000000	0.0000000	0.0000000
sociable	1	0.1516940	0.1408590	0.0000000	0.0000000
clínica	5	0.1516940	0.0821678	0.0000000	0.0821678
Totales		0.7487777	0.4117010	0.1041784	0.2453934

Word 3	sociable	Sences: 1
		0
dejar	14	0.0376533
gato	7	0.0428550
normalmente	0	0.0000000
pensión	4	0.0318069
clínica	5	0.0785674
Totales		0.1908826

Word 5	clínica	Sences: 5				
		0	1	2	3	4
dejar	14	0.0921285	0.0240335	0.0000000	0.0921285	0.0921285
gato	7	0.1242260	0.0338798	0.0000000	0.1242260	0.1242260
normalmente	0	0.1443380	0.0000000	0.0000000	0.1443380	0.1443380
sociable	1	0.1571350	0.2357020	0.0000000	0.1571350	0.1571350
pensión	4	0.1961420	0.0954206	0.0000000	0.1643360	0.1643360
Totales		0.7139695	0.3890359	0.0000000	0.6821635	0.6821635

3.3.2 Ejemplo de mediciones del contexto 2.

3.3.2.1 Contexto 2.Método de Lesk original.

Contexto "*Detienen a militar implicado en el asesinato de policías*"

Contexto
Detienen a militar implicado en el asesinato de policías

# words	6	window sz	5

Word 0	detienen	Sences: 5				
		0	1	2	3	4
militar	4	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000	0.0000000	0.1178510	0.0000000
el	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
policías	2	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.1178510	0.0000000

Word 3	el	Sences: 0
detienen	5	
militar	4	
implicado	2	
asesinato	1	
policías	2	
Totales		

Word 1	militar	Sences: 4			
		0	1	2	3
detienen	5	0.0000000	0.0000000	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000	0.0000000	0.0000000
el	0	0.0000000	0.0000000	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000	0.0000000	0.0000000
policías	2	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000

Word 4	asesinato	Sences: 1
		0
detienen	5	0.0000000
militar	4	0.0000000
implicado	2	0.0000000
el	0	0.0000000
policías	2	0.0000000
Totales		0.0000000

Word 2	implicado	Sences: 2	
		0	1
detienen	5	0.0785674	0.0000000
militar	4	0.0000000	0.0000000
el	0	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000
policías	2	0.0000000	0.0000000
Totales		0.0785674	0.0000000

Word 5	policías	Sences: 2	
		0	1
detienen	5	0.0000000	0.0000000
militar	4	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000
el	0	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000
Totales		0.0000000	0.0000000

3.3.2.2 Contexto 2. Método de Lesk Modificado.

Contexto "*Detienen a militar implicado en el asesinato de policías*"

Contexto
Detienen a militar implicado en el asesinato de policías

# words	6	window sz	5

Word 0	detienen	Sences: 5				
		0	1	2	3	4
militar	4	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
el	0	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
policías	2	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

Word 3	el	Sences: 0
detienen	5	
militar	4	
implicado	2	
asesinato	1	
policías	2	
Totales		

Word 1	militar	Sences: 4			
		0	1	2	3
detienen	5	0.0000000	0.0000000	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000	0.0000000	0.0000000
el	0	0.0000000	0.0000000	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000	0.0000000	0.0000000
policías	2	0.0000000	0.0000000	0.0000000	0.0000000
Totales		0.0000000	0.0000000	0.0000000	0.0000000

Word 4	asesinato	Sences: 1
		0
detienen	5	0.0000000
militar	4	0.0000000
implicado	2	0.0000000
el	0	0.0000000
policías	2	0.0000000
Totales		0.0000000

Word 2	implicado	Sences: 2	
		0	1
detienen	5	0.0000000	0.0000000
militar	4	0.0000000	0.0000000
el	0	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000
policías	2	0.0000000	0.0000000
Totales		0.0000000	0.0000000

Word 5	policías	Sences: 2	
		0	1
detienen	5	0.0000000	0.0000000
militar	4	0.0000000	0.0000000
implicado	2	0.0000000	0.0000000
el	0	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000
Totales		0.0000000	0.0000000

3.3.2.3 Contexto 2. Método Propuesto.

Contexto "*Detienen a militar implicado en el asesinato de policías*"

Contexto
Detienen a militar implicado en el asesinato de policías

# words	6	window sz	5

Word 0	detienen	Sences: 5				
		0	1	2	3	4
militar	4	0.2730500	0.1972030	0.0000000	0.1972030	0.1972030
implicado	2	0.1885620	0.1885620	0.0000000	0.4242640	0.3064130
el	0	0.1732050	0.1732050	0.0000000	0.1732050	0.1732050
asesinato	1	0.3975230	0.1490710	0.0000000	0.1490710	0.1490710
policías	2	0.1895210	0.1105540	0.0000000	0.1105540	0.1105540
Totales		1.2218610	0.8185950	0.0000000	1.0542970	0.9364460

Word 3	el	Sences: 0
detienen	5	
militar	4	
implicado	2	
asesinato	1	
policías	2	
Totales		

Word 1	militar	Sences: 4			
		0	1	2	3
detienen	5	0.0821678	0.1972030	0.0000000	0.1972030
implicado	2	0.0000000	0.3204540	0.1232520	0.3204540
el	0	0.0000000	0.1885620	0.0000000	0.1885620
asesinato	1	0.2886750	0.1732050	0.1237180	0.1732050
policías	2	0.0532397	0.3620300	0.1064790	0.2555510
Totales		0.4240825	1.2414540	0.3534490	1.1349750

Word 4	asesinato	Sences: 1
		0
detienen	5	0.1242260
militar	4	0.1332350
implicado	2	0.0000000
el	0	0.0000000
policías	2	0.0000000
Totales		0.2574610

Word 2	implicado	Sences: 2	
		0	1
detienen	5	0.2671290	0.3456970
militar	4	0.2730500	0.2730500
el	0	0.1972030	0.1972030
asesinato	1	0.1885620	0.1885620
policías	2	0.1732050	0.2350640
Totales		1.0991490	1.2395760

Word 5	policías	Sences: 2	
		0	1
detienen	5	0.0460642	0.0000000
militar	4	0.3440100	0.0000000
implicado	2	0.0721688	0.0000000
el	0	0.0000000	0.0000000
asesinato	1	0.0000000	0.0000000
Totales		0.4622430	0.0000000

3.4 Comparación de los resultados

3.4.1 Contexto 1.

Contexto	1
Dejar a un gato normalmente sociable en una pensión o en la clínica.	

Palabras		**Cant. Sent.**	**Correcto**	**Lesk1**		**Lesk2**		**Ledo**	
1	Dejar	14	0	1, 4 y 6	✘	Todos	✘	0	✓
2	Gato	7	0 y 2	4 y 5	✘	Todos	✘	4 y 6	✘
3	Normalmente	0	Ninguno	Ninguno	✓	Ninguno	✓	Ninguno	✓
4	Sociable	1	0	0	✓	0	✓	0	✓
5	Pensión	4	0	1	✘	Todos	✘	0	✓
6	Clínica	5	0	Todos	✘	3 y 4	✘	0	✓
Aciertos correctos			6	2		2		**5**	
% de aciertos			100%	33.33%		33.33%		**83.33%**	

3.4.2 Contexto 2.

Contexto	2
Detienen a militar implicado en el asesinato de policías	

Palabras		**Cant. Sent.**	**Correcto**	**Lesk1**		**Lesk2**		**Ledo**	
1	Detienen	5	1	3	✘	Todos	✘	0	✘
2	Militar	4	0, 1 y 3	Todos	✘	Todos	✘	1	✓
3	implicado	2	0	0	✓	Todos	✘	1	✘
4	El	0	ninguno	ninguno	✓	ninguno	✓	ninguno	✓
5	asesinato	1	0	0	✓	0	✓	0	✓
6	Policías	2	0	Todos	✘	Todos	✘	0	✓
Aciertos correctos			6	3		2		**4**	
% de aciertos			100%	50.00%		33.33%		**66.66%**	

3.4.3 Resultados de los experimentos

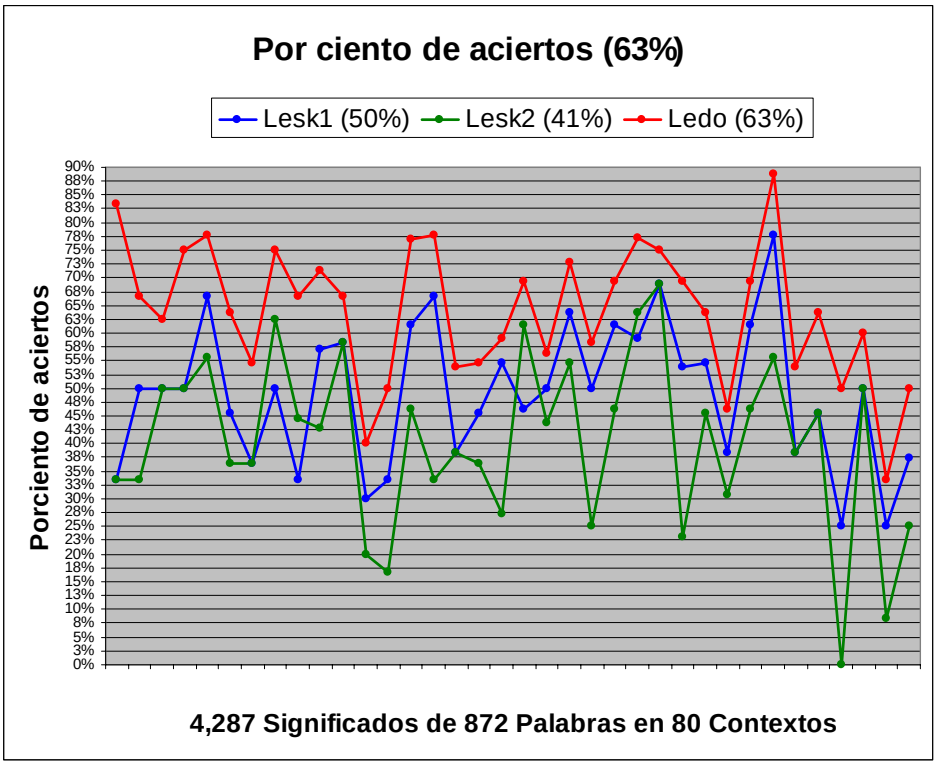
3.4.3.1 Listado de por cientos de aciertos

El método propuesto obtuvo un 63% de aciertos siendo un **13% mejor** que el Método de Lesk original (50%) y un **22% mejor** que el Método de Lesk Modificado (41%).

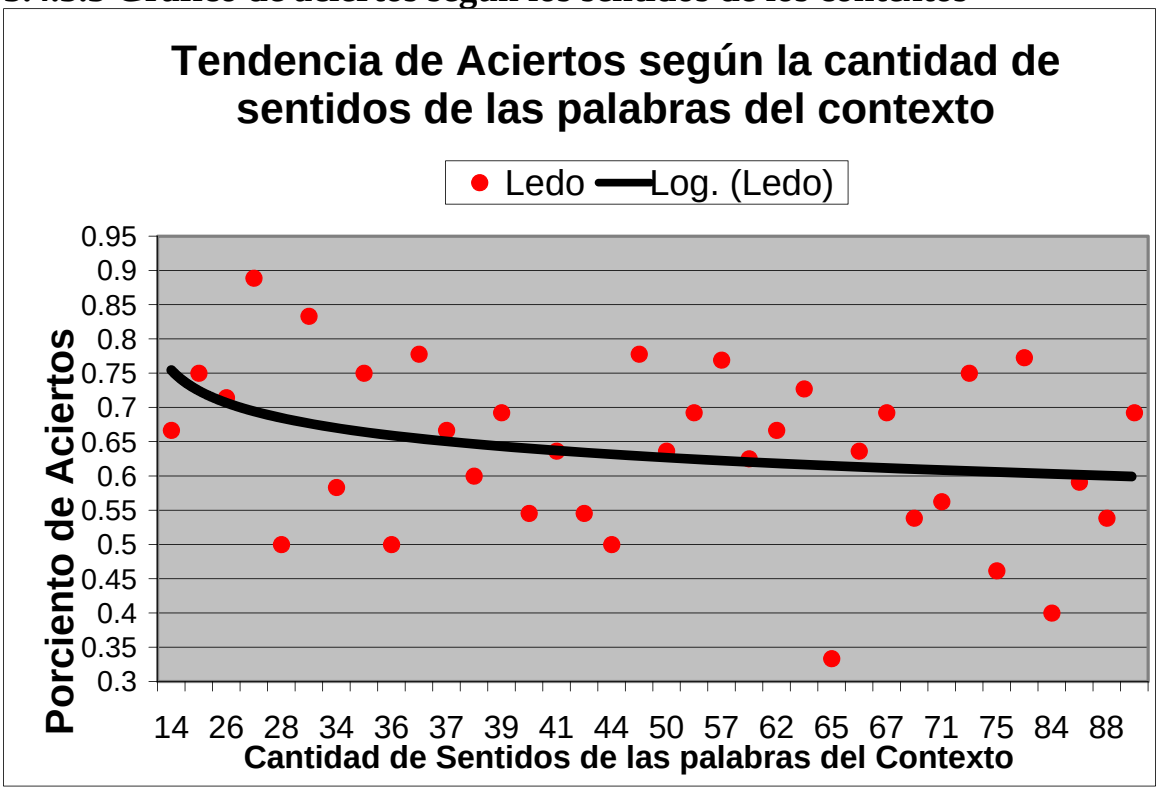
Contexto			Lesk1		Lesk2		Ledo	
No	**Palabras**	**Sentidos**	Aciertos	%	Aciertos	%	Aciertos	%
1.	6	31	2	33%	2	33%	5	83%
2.	6	14	3	50%	2	33%	4	67%
3.	9	50	3	33%	2	22%	3	33%
4.	5	12	2	40%	1	20%	2	40%
5.	9	33	7	78%	4	44%	5	56%
6.	9	34	6	67%	3	33%	4	44%
7.	8	59	4	50%	4	50%	5	63%
8.	8	21	4	50%	4	50%	6	75%
9.	9	37	6	67%	5	56%	7	78%
10.	11	66	5	45%	4	36%	7	64%
11.	10	32	7	70%	4	40%	5	50%
12.	8	41	2	25%	3	38%	2	25%
13.	10	45	6	60%	5	50%	5	50%
14.	8	47	2	25%	0	0%	1	13%
15.	11	40	4	36%	4	36%	6	55%
16.	8	36	4	50%	5	63%	6	75%
17.	12	57	8	67%	5	42%	8	67%
18.	5	32	3	60%	3	60%	3	60%
19.	9	37	3	33%	4	44%	6	67%
20.	7	26	4	57%	3	43%	5	71%
21.	6	18	1	17%	1	17%	1	17%
22.	8	50	2	25%	3	38%	2	25%
23.	7	47	3	43%	3	43%	3	43%
24.	12	62	7	58%	7	58%	8	67%
25.	7	29	2	29%	2	29%	2	29%
26.	13	67	10	77%	7	54%	10	77%
27.	10	84	3	30%	2	20%	4	40%
28.	13	172	5	38%	4	31%	4	31%
29.	7	65	1	14%	1	14%	1	14%
30.	10	38	7	70%	5	50%	7	70%
31.	11	75	5	45%	2	18%	4	36%
32.	6	36	2	33%	1	17%	3	50%
33.	6	58	4	67%	2	33%	3	50%
34.	13	57	8	62%	6	46%	10	77%
35.	9	50	6	67%	3	33%	7	78%
36.	13	88	5	38%	5	38%	7	54%

37.	11	42	5	45%	4	36%	6	55%
38.	5	10	2	40%	3	60%	3	60%
39.	9	23	7	78%	3	33%	5	56%
40.	12	48	7	58%	6	50%	6	50%
41.	10	38	5	50%	3	30%	4	40%
42.	10	31	7	70%	3	30%	5	50%
43.	22	87	12	55%	6	27%	13	59%
44.	13	67	6	46%	8	62%	9	69%
45.	15	67	8	53%	7	47%	8	53%
46.	16	71	8	50%	7	44%	9	56%
47.	11	65	7	64%	6	55%	8	73%
48.	12	34	6	50%	3	25%	7	58%
49.	13	39	8	62%	6	46%	9	69%
50.	26	123	13	50%	14	54%	14	54%
51.	17	101	13	76%	6	35%	13	76%
52.	10	26	6	60%	5	50%	6	60%
53.	18	70	9	50%	7	39%	9	50%
54.	22	82	13	59%	14	64%	17	77%
55.	13	39	10	77%	8	62%	10	77%
56.	16	74	11	69%	11	69%	12	75%
57.	13	72	6	46%	4	31%	6	46%
58.	21	102	13	62%	9	43%	11	52%
59.	13	91	7	54%	3	23%	9	69%
60.	11	41	6	55%	5	45%	7	64%
61.	14	55	10	71%	7	50%	9	64%
62.	12	70	4	33%	3	25%	4	33%
63.	13	75	5	38%	4	31%	6	46%
64.	13	55	8	62%	6	46%	9	69%
65.	9	26	7	78%	5	56%	8	89%
66.	17	111	8	47%	8	47%	8	47%
67.	13	46	7	54%	7	54%	7	54%
68.	13	67	5	38%	5	38%	7	54%
69.	12	66	4	33%	2	17%	4	33%
70.	11	50	5	45%	5	45%	7	64%
71.	4	28	1	25%	0	0%	2	50%
72.	10	68	7	70%	4	40%	4	40%
73.	7	46	5	71%	5	71%	5	71%
74.	10	37	5	50%	5	50%	6	60%
75.	10	42	7	70%	3	30%	5	50%
76.	7	46	2	29%	1	14%	2	29%
77.	9	63	3	33%	5	56%	5	56%
78.	12	65	3	25%	1	8%	4	33%
79.	8	44	3	38%	2	25%	4	50%
80.	10	38	6	60%	6	60%	6	60%
Totales	**872**	**4,287**	**436**	**50%**	**358**	**41%**	**549**	**63%**

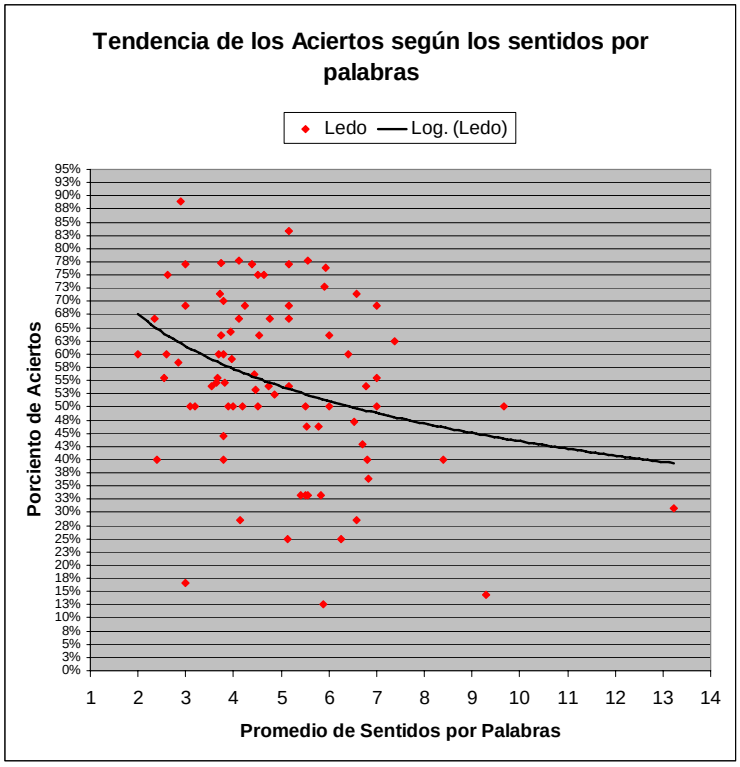
3.4.3.2 Gráfico con muestras de aciertos



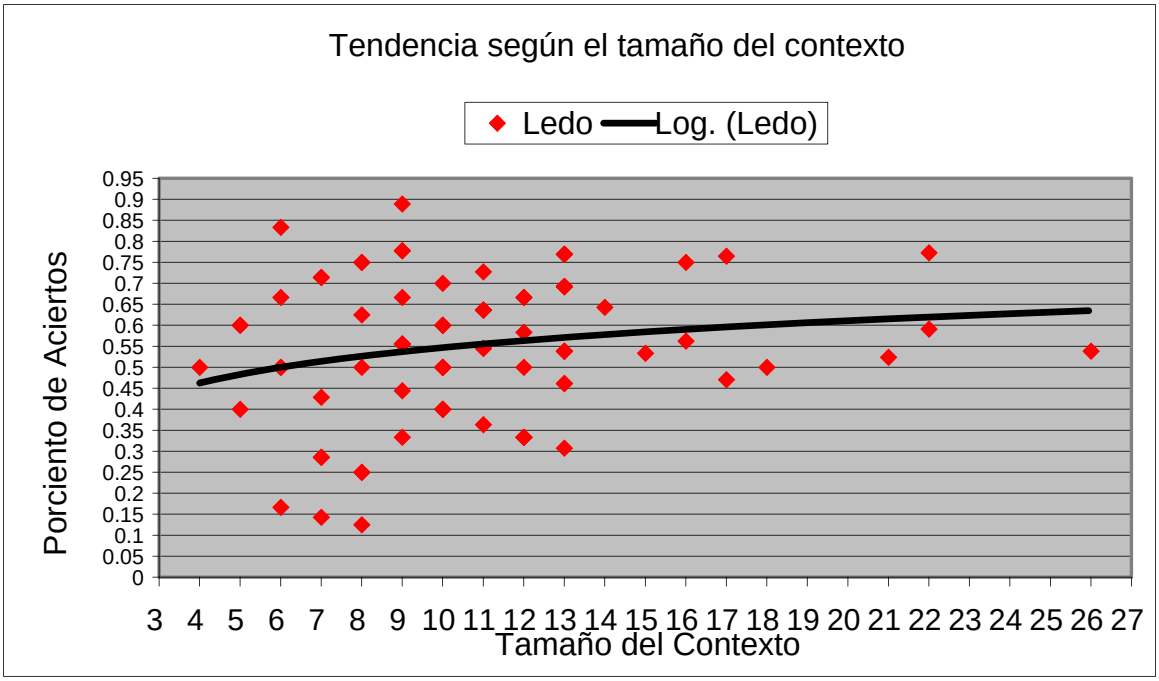
3.4.3.3 Gráfico de aciertos según los sentidos de los contextos



3.4.3.4 Gráfico de aciertos según los sentidos de las palabras

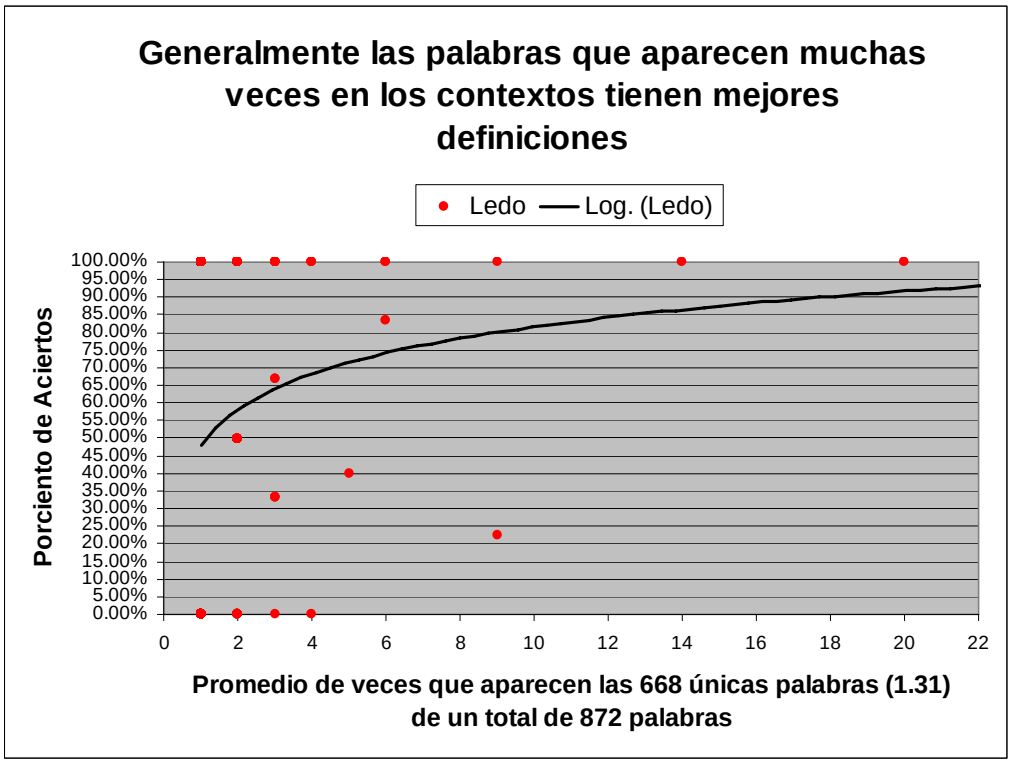


3.4.3.5 Gráfico de aciertos según el tamaño del contexto



3.4.3.5 Gráfico de aciertos según los significados de únicas palabras

3.4.3.6 Gráfico de la calidad de los significados



3.5 Méritos y difusión durante el período del Doctorado.

3.5.1 Resumen.

- Presea “Lázaro Cárdenas” 2OO3.
- Mejor estudiante de Doctorado del IPN 2003
- Mejor estudiante del Doctorado en Ciencias de la Computación del CIC 2003.
- 10.00 de promedio de calificaciones.
- 18 Publicaciones en revistas y libros.
- 28 Publicaciones y participaciones en eventos internacionales.
- 11 Publicaciones y participaciones en eventos nacionales.
- 4 Proyectos de investigación-desarrollo.
- 32 Cursos de licenciatura impartidos.
- 3 Pláticas impartidas.
- 5 Participaciones en seminarios de investigación
- 3 Tesis de Maestría asesoradas
- 7 Tesis de Licenciatura asesoradas
- Otras actividades científicas y académicas.

3.5.2 Publicaciones más importantes realizadas.

1. Y. Ledo-Mezquita y G. Sidorov. ***Combinación de los métodos de Lesk original y simplificado para desambiguación de sentidos de palabras***. In: Alexander Gelbukh, Manuel Montes y Gómez (Eds.). Advances in Natural Language Understanding and Intelligent Access to Textual Information. Proceedings of First International Workshop on Natural Language Understanding and Intelligent Access to Textual Information, in conjunction with MICAI-2005, November 15, 2005, Monterrey, Mexico, 2005; part of: Hugo Terashima-Marín, Horacio Martínez-Alfaro, Manuel Valenzuela-Rendón, Ramón Brena-Pinero (Eds.). Tutorials and Workshops Proceedings of Fourth Mexican International Conference on Artificial Intelligence, ISBN: 968-891-094-5, pp. 41-47.
2. Alexander Gelbukh, Grigori Sidorov, Yoel Ledo-Mezquita. ***Some Linguistic Methods of Improving the Quality of Document Retrieval on the Internet***. International Journal of Electronic Business (IJEB), ISSN 1470-6067, ISSN 1741-5063, Vol. 3, No. 3, May–June 2005, Special Issue “Multidisciplinary, Interdisciplinary, and Transdisciplinary Research in Electronic Business,” Part I.
3. Alexander Gelbukh, Grigori Sidorov, Yoel Ledo-Mezquita. ***On similarity of word senses in explanatory dictionaries***. International Journal of Translation, BAHRI Publications, New Delhi, India. ISSN 0970-9819, Vol.15, No. 2, 2003, pp. 51–60.
4. Yoel Ledo Mezquita, Grigori Sidorov, Alexander Gelbukh. ***Tool for Computer-Aided Spanish Word Sense Disambiguation***. In: Computational Linguistics and Intelligent Text Processing (CICLing-2003, Mexico City). Lecture Notes in Computer Science

(indexed by SCIE), N 2588, ISSN 0302-9743, ISBN 3-540-00532-3, Springer-Verlag, pp. 277–280.

5. C. Anías y Y. Ledo (Eds.) ***La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad*** Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
6. Yoel Ledo Mezquita, Grigori Sidorov, Alexander Gelbukh, Caridad Anías Calderón. ***Búsqueda en bibliotecas digitales indexadas con sentidos de palabras*** tomo I, pp288-299, en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
7. Yoel Ledo Mezquita, Julio Cesar Jerez Camps, Yordanys Pantoja Parra ***Sistema de Mensajería Unificada para una Intranet*** Tomo I, pp313-316, en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
8. Yoel Ledo Mezquita, Julio Cesar Camps, Ramsés Felipe González, Yordanys J. Pantoja Parra. ***Diseminación del conocimiento multimedia en Intranets e Internet*** Tomo II, pp426-433, en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
9. Yoel Ledo Mezquita, Grigori Sidorov, Alexander Gelbukh. ***Sistema de asignación semiautomático de sentidos de palabras en textos en español*** Tomo II, pp625-631, en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
10. Yoel Ledo Mezquita, Rubén Díaz Zamora, Yurdik Cervantes Mendoza. ***Gestión dinámica de noticias*** Tomo II, pp639-646 en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
11. Yoel Ledo Mezquita, Reinaldo Díaz Castro, Yordanys Pantoja Parra. ***Recuperación de información en Internet*** Tomo II, pp690-695 en el Libro "La Telemática y su aplicación en la educación a distancia y en la informatización de la sociedad" C. Anías y Y. Ledo (Eds.) Editorial Universitaria, La Habana, Cuba, 2002, ISBN 959-261-042-8
12. Yoel Ledo Mezquita, Grigori Sidorov, Alexander Gelbukh. ***Herramienta para la desambiguación semiautomática de sentidos de palabras en Español.*** J.U. Sossa Azuela *et al*. (Eds.) Avances en Ciencias de la Computación e Ingeniería de Cómputo. Proc. CIC'2002, XI Congreso Internacional de Computación, November 2002, CIC-IPN, ISBN 970-18-8590-2, ISBN 970-18-8591-0, Mexico, v. II, p. 343–347.
13. Yoel Ledo Mezquita, Grigori Sidorov, Alexander Gelbukh, and Caridad Anías Calderón. ***Búsqueda inteligente en bibliotecas digitales***. In: Tópicos Avanzados de Minería de Datos y Sistemas de Información, Proc. of 3^{r} Taller Internacional de Sistemas de Información and 4^{o} Taller Internacional de Minería de Datos, CIC, IPN, October 1–3, 2002. Vol. 1., CD edition: ISBN 970-18-8545-7; complete book: ISBN 970-18-8546-5.
14. Yoel Ledo, Damian Bergantiños, Alejandro Machado, William Azcuy, Yadenis Piñero, Camilo Nuñez, ***Desarrollo de Portales Dinámicos con Procesamientode***

Texto. Experiencia Universitaria, Serie Roja, No. 98. Centro de Investigación en Computación, Instituto Politénico Nacional, 1998, ISBN 970-18-7174-X.

Conclusiones

En estas conclusiones se resumen los resultados de la investigación de la tesis doctoral

Conclusiones

Resultados, aportes y contribuciones

Método propuesto

El método propone la desambiguación de los sentidos de las palabras del contexto basado en la comparación de los sentidos de la palabra analizada en relación al contexto y a los sentidos de las palabras que conforman el contexto, teniendo en cuenta la influencia de cada palabra del contexto según la distancia a la que se encuentra de la palabra analizada y la influencia de la semejanza en función de muchos recursos léxicos.

Los resultados obtenidos demuestran que el método obtuvo mejor precisión que los otros dos métodos con los cuales se comparó.

Métodos	**Lesk 1**	**Lesk 2**	**Ledo**
Contextos	80	80	80
Significados	4287	4287	4287
Palabras	872	872	872
Aciertos	436	358	549
Fallos	436	514	323
% de Acierto	**50%**	**41%**	**63%**

Semejanza o similitud entre dos textos

El método propone una nueva forma de determinar la semejanza usando diferentes recursos léxicos tales como el diccionario explicativo con definiciones normalizadas, sinónimos, antónimos, merónimos (parte de), holónimo (contiene a), hipónimos, hiperónimos en el cual cada recurso aporta a la determinación de la semejanza obteniéndose mejores resultados con valores más discretos normalizados entre 0 y 1.

Preparación y conversión de los recursos léxicos

El método usa algunos recursos léxicos que ya existen en la actualidad, los cuales fueron convertidos a bases de datos normalizadas e indexadas, de forma que permite un procesamiento acelerado de los análisis realizados en tiempos muy cortos desde el punto de vista de la persona que espera por los resultados.

El corpus

Se creó un corpus con 100 documentos para evaluar los WSD para el español, de este corpus se seleccionaron los contextos usados para hacer las mediciones para los distintos métodos de desambiguación.

Distancia y tamaño del contexto

El método propone una atenuación de la influencia en el peso de las palabras según la distancia a la que se encuentren de la palabra analizada, de forma tal, que palabras más cercanas tienen mayor influencia y que palabras más lejanas tienen menor influencia, expresándose de forma discreta bajo una curva exponencial con tendencia a cero hacia el infinito.

El español en el análisis

El método fue diseñado 100% para ser usado en textos en español, pues existen pocas investigaciones en el tema de la desambiguación en comparación con otros idiomas, específicamente con el inglés.

Uso práctico

Disponer de un método de desambiguación de sentidos de palabras para el idioma español para la recuperación inteligente de información en buscadores de Internet ya que en la actualidad tanto el volumen existente como la demanda en su consulta son grandes.

Rumbos de investigaciones posteriores

Aplicar algoritmos genéticos para la determinación del peso total combinatorio que existen entre todas las palabras del contexto de forma tal que se determine la mejor combinación de significados para todas las palabras al mismo tiempo y no el mejor peso para cada palabra de forma independiente.

Glosario

En este glosario se presentan las definiciones de términos, palabras, siglas, etcétera, que pueden contribuir al entendimiento del documento.

Glosario

Conocimientos Representación simbólica de ideas, conceptos, nociones, hechos, seres, acciones, y de las relaciones entre ese tipo de elementos que reflejan un dominio del universo físico o del mundo de las ideas. [Inglés: knowledge]

Desambiguación Eliminación de ambigüedades. [inglés: disambiguation]

Diacrítico Se dice de una marca pequeña que se usa con una vocal para indicar o una variación en el mismo sonido (por ejemplo: el acento o un tono particular) o un sonido diferente, aunque muchas veces parecido. Los diacríticos (o marcas diacríticas) comunes incluyen el acento agudo (◌́), el acento grave (◌̀), el macrón (◌̄), el circunflejo (◌̂), el circunflejo inverso ("cuernitos",◌̌), la tilde (◌̃), y la diéresis (◌̈). [Inglés: diacritic]

Enunciado [1]: término a veces usado para significar oración (oración independiente). [Inglés: sentence]

[2]: secuencia completa de palabras dichas (o potencialmente dichas) por un hablante. A veces contiene uno o más enunciados en el sentido [1], pero a veces no. Por ejemplo: las palabras "*como usted mande*" suelen formar un enunciado en este sentido, pero no forman una oración independiente. [Inglés: utterance]

Fonema, fonémico Un fonema es un sonido que contrasta con otros sonidos de una lengua. Las distinciones entre fonemas (es decir, lo que causa contraste) son diferencias fonémicas.

Dos sonidos que tienen una diferencia en su pronunciación difieren fonéticamente, pero si esa diferencia no produce contraste, no es fonémica sino alofónica. Por ejemplo: en español la n de "*mando*" y la de "*mango*" difieren fonéticamente (la primera es apical [n] y la otra es velar [ŋ]). Pero esa diferencia no es fonémica: en

el sistema del español el contraste es llevado por la consonante siguiente (d vs. g), y la **n** varía su pronunciación según la articulación de esa consonante. Así, decimos que hay un solo fonéma /n/, el cual tiene dos alófonos, [n] y [ŋ], (a veces los fonemas se encierran entre diagonales.) Los pares mínimos son especialmente importantes para establecer que una distinción es fonémica. En general, cada fonema debe representarse en una ortografía práctica mediante un grafema distinto (por ejemplo: una letra del alfabeto), pero las diferencias entre sus alófonos no deben representarse. [Inglés: phoneme, phonemic]

Fonética, fonético La fonética es el estudio de los sonidos del lenguaje desde un punto de vista (relativamente) detallado y objetivo, no tomando en cuenta su función en el sistema lingüístico de la lengua. Es la parte de la fonología que trata de la manera en que se pronuncian los sonidos y su forma acústica. A menudo lo fonético se encierra entre corchetes, por ejemplo: [n] indica "el sonido fonético 'n', sea fonema o no". [Inglés: phonetics, phonetic]

Fonología La fonología es el estudio de los sonidos del lenguaje. La fonética es la parte de la fonología que trata de la manera en que se pronuncian los sonidos y su forma acústica, pero la fonología también incluye el estudio de la manera en que los sonidos funcionan sistemáticamente en la lengua. Las otras divisiones importantes de ese estudio incluyen el análisis de los fonemas, los cambios de un sonido a otro en ciertos contextos (la alternancia morfofonémica y alofónica), la estructura de las sílabas, el acento, la entonación, y el tono lingüístico. [Inglés: phonology]

Frase Grupo de una o más palabras que funciona como unidad, pero que (normalmente) no funciona en su totalidad independientemente, como en el caso de una oración. A veces se marcan las frases, como las oraciones, poniéndolas entre corchetes. Por ejemplo: [*las montañas* [*más altas*]] es una frase nominal, y también contiene una frase adjetival, [*más altas*]. Contrástese con oración. [Inglés: phrase]

Gramática Es la manera característica en que se combinan los elementos básicos (especialmente los elementos léxicos) de una lengua, para formar estructuras más complejas que permitan la comunicación de los pensamientos. La gramática incluye la morfología y la sintaxis; algunos analistas incluyen la fonología, la semántica y el léxico también como parte de la gramática. [Inglés: grammar]

Hipónimo, Hiperónimo

La semántica denomina hipónimo a aquella palabra que posee todos los rasgos semánticos, o semas, de otra más general, su hiperónimo, pero que añade en su definición otros rasgos semánticos que la diferencian de la segunda. Por ejemplo, descapotable comparte con coche todos sus rasgos mínimos, a saber [+vehículo], [+con motor], [+pequeño tamaño], etc., pero añade a éstos el rasgo [+sin capota].

La semántica denomina hiperónimo a aquel término general que puede ser utilizado para referirse a la realidad nombrada por un término más particular. Semánticamente, un hiperónimo no posee ningún rasgo semántico, o sema, que no comparta su hipónimo, mientras que éste sí posee rasgos semánticos que lo diferencian de aquél. Por ejemplo, coche posee sólo los semas [+vehículo], [+con motor] y [+pequeño tamaño], que comparte con descapotable, mientras que descapotable posee además el rasgo [+sin capota], que lo diferencia de [+coche]. Al redactar un texto conviene utilizar hiperónimos para evitar la repetición de palabras ya empleadas anteriormente, como se hace en el siguiente ejemplo: *De repente, un descapotable rojo paró frente al banco. Del automóvil salieron dos individuos encapuchados, mientras otro esperaba en el vehículo.*

Inteligencia Artificial (IA)

Disciplina dedicada a desarrollar y aplicar enfoques computacionales al comportamiento inteligente. Estudia preferentemente los comportamientos y fenómenos de percepción, solución de problemas, razonamiento, utilización de un lenguaje natural y planeamiento de actividades. [Inglés: Artificial Intelligence (AI)]

Lenguaje natural (LN)

Manipulación de expresiones de un idioma humano que permite a un sistema computacional obedecer comandos en ese lenguaje y/o entregar resultados en él, permitiendo un manejo del lenguaje con una libertad comparable a la que maneja un ser humano típico (no en estructuras rígidas y muy limitadas). [Inglés: Natural Language (NL)]

Lemma, Lematización

Término en Latin. Forma canónica. En lexicografía, entrada léxica del diccionario en que se suministra diversa información y es a menudo representativa de distintas formas flexionadas; por ejemplo: ir es el lema de voy, vas, íbamos, fueron y el resto de sus formas conjugadas. Lematización.

En lexicografía se denomina ‘lematización’ al proceso de reducción de las diferentes formas flexivas de una palabra a la forma canónica que se selecciona como lema. [Inglés: lemma]

Lexema Unidad léxica abstracta que no puede descomponerse en otras menores, aunque sí combinarse con otras para formar compuestos, y que posee un significado definible por el diccionario, no por la gramática. Por ejemplo: *fácil* es el lexema básico de *facilidad*, *facilitar*, *fácilmente*. [Inglés: lexeme]

Léxico El conjunto de los morfemas de una lengua, junto con raíces complejas o palabras pre-formadas (o sea que no se arman en forma productiva), modismos y otras frases establecidas. Éstas son las estructuras lingüísticas que un hablante sabe como unidades completas y que puede usar sin tener que determinar sus significados a base de sus partes integrantes. Un diccionario (que a veces también se llama léxico) es un libro que exhibe elementos del léxico de una lengua, especialmente palabras, con una indicación breve de sus significados y usos. Contrástese con gramática; sintaxis, morfología, fonología, semántica. [Inglés: lexicon]

Merónimo, Holónimo Meronimia es una relación semántica. Un merónimo es el nombre atribuido a un constituyente que forma parte de, que es substancia de o que es miembro de algo. Meronimia es lo opuesto a la holonimia.

Morfema Forma mínima con significado, es decir, una forma que tiene significado pero que no puede dividirse en otras formas más pequeñas que tienen significado. Las dos clases mayores de morfemas son las raíces y los afijos. Por ejemplo: en la palabra "*queríamos*" se pueden reconocer tres morfemas: la raíz "*quer*" 'querer', el sufijo "*-ía*" 'copretérito', y el sufijo "*-mos*" '1ª persona de plural'. Véase también morfema cero, palabra. [Inglés: morpheme]

Morfología La morfología es el estudio de cómo los morfemas se combinan para formar raíces complejas y palabras. Compárese con gramática; contrástese con sintaxis, fonología, semántica, léxico. [Inglés: morphology]

Oración [1]: frase verbal junto con las frases nominales o adverbiales u otras oraciones que dependan de ella. Las oraciones pueden ser dependientes o independientes. Por ejemplo, la oración independiente "*Dice Juan que te buscaba*" contiene la

oración dependiente "*que te buscaba*". A veces se marcan las oraciones poniéndolas entre corchetes. El estudio de la estructura de las oraciones es uno de los temas centrales de la sintaxis. [Inglés: clause]

[2]: a veces se usa en contraste con el término cláusula para indicar una oración independiente, con las cláusulas dependientes que pueda tener, o una serie de dos o más oraciones independientes coordenadas con una conjunción como "*y*" u "*o*". Compárese con enunciado. [inglés: sentence]

Palabra Es una raíz, junto con los afijos que dependan de ella y posiblemente de otras raíces (en el caso de una raíz compuesta), que puede pronunciarse sola en el uso normal de una lengua, por ejemplo, como respuesta a una pregunta. Frecuentemente las palabras tienen rasgos fonológicos especiales. En el náhuatl de Tetelcingo, por ejemplo, las palabras normalmente se pueden reconocer por el penúltimo acento. Compárese con frase, morfema. [Inglés: word]

Paradigma Lista de formas relacionadas de una palabra, especialmente si son relacionadas por flexión. Por ejemplo: "*hablo, hablas, habla*" es un paradigma de formas de tiempo presente singular del verbo "*hablar*" en el español; "*hablar, hablante, hablado*" es un paradigma de formas infinitiva del mismo verbo. La yuxtaposición de paradigmas paralelos en forma de un cuadro, puede facilitar la comparación y por lo tanto el análisis de las formas. [Inglés: paradigm]

Semántica La Semántica es el estudio de los significados de las estructuras de las lenguas (morfemas, palabras, frases, oraciones y otras). Una diferencia o semejanza semántica es una diferencia o semejanza de significados. Compárese con gramática, sintaxis, morfología, fonología, léxico. [Inglés: semantic, semantic]

Sintaxis Es el estudio de cómo las palabras se combinan para formar frases y oraciones, ya sean dependientes o independientes. Compárese con gramática; contrástese con morfología, fonología, semántica, léxico. [Inglés: syntax]

Referencias

En las referencias se presenta la bibliografía consultada y referenciada enumerada y ordenada alfabéticamente.

Referencias

[1] **Aguirre E. and Rigau G. (1996).** Word Sense Disambiguation using Conceptual Density. Proc. 16th international conference on COLING. Copenhangen.

[2] **Alexandrov, M; Gelbukh, A.; y Makaganov, P. (2000).** On Metrics for Keyword-based Document Selectión and Classification. In Conference on Intelligent Text Processing and Computational Linguistics CICLing-2000. Centro de Investigación en Computación. Instituto Politécnico Nacional. February 13-19, 2000. México City, México. 373-389.

[3] **Alshawi, Hiyan y Carter, David (1994).** Training and scaling preference functions for disambiguation. Computacional Linguistics, 20(4), 635-648.

[4] **Anthony, Edward (1954).** An exploratory inquiry into lexical clusters. American Speech, 29(3). 175-180.

[5] **Aone, C. y McKee, D. (1993).** A language-independent anaphora resolution system for understanding multilingual texts. In: Proceedings of the 31st Annual Meeting of the ACL (ACL'93). The Ohio State University, Columbus, Ohio., 156-163.

[6] **Asprejan, Jurij D. (1974).** Regular polysemy. Linguistics, 142, 5-32.

[7] **Atkins, Beryl T. S. (1987).** Semantic ID tags: corpus evidence for dictionary senses. Proceedings of the Third Annual Conference of the UW Center for the New OED, Waterloo, Ontario, 17-36.

[8] **Avancini, Henri, (2000).** Information Retrieval & Automated Web Page Categorization. ISISTAN – Research Report RRxx-00. Tandil (Bs. As.).

[9] **Ayto, John R. (1983)**. On specifying meaning. In Hartmann, R.R.K. (Ed.), Lexicography: Principles and Practice, Academic Press, London, 89-98.

[10] **Baeza, Ricardo y Ribeiro, Bertier (1999).** Moderm Information Retrieval. Adison Wesley, ACM Press, New Cork.

[11] **Berry-Rogghe, Godelieve (1973).** The computation of collocations and their relevance to lexical studies. In Aitken, Adam J.; Bailey, Richard W., and Hamilton- Smith, Neil (Eds.) The Computer and Literary Studies. Edinburgh University Press, Edinburgh, United Kingdom, 103-112.

[12] **Bloomfield, Leonard (1933).** Language. Holt, New York.

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

[14] **Bruce, Rebecca y Wiebe, Janyce (1994).** Word-sense Disambiguation Using Decomposable Models. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 139-145.

[15] **Campos, Luis M de. (2001).** Un modelo de recuperación de información basado en redes bayesianas. Universidad de Granada, España.

[16] **Connine, Cynthia (1990).** Effects of sentence context and lexical knowledge in speech processing. In Altmann, Gerry T. (Ed.) Cognitive models in speech processing. The MIT Press. Cambridge, Massachussets, 540pp.

[17] **Carter, D. (1987)** Interpreting Anaphora in Natural Language Texts. Ellis Horwood, Chichester

[18] **Cerdá, Massó Ramón (1975)** Lingüística Hoy. Colección “Hay que saber”. 3ª Edición, Teide. Barcelona, España.

[19] **Cowie J., Guthrie J. and Guthrie L. (1992)** Lexical disambiguation using simulated annealinf. Proc. DARPA Workshop on Speech an Natural Language. 238-242. New York.

[20] **Cruse, D. A. (1986).** Lexical Semantics. Cambridge University Press. Cambridge, United Kingdom.

[21] **Dahlgren, Kathleen G. (1988).** Naive Semantics for Natural Language Understanding. Kluwer Academic Publishers, Boston. 258pp.

[22] **Dolan, William; Vanderwende, Lucy; y Richardson, Stephen (2000).** Polysemy in a Broad-Coverage Natural Language Processing System. In Polysemy: Theoretical and Computational Approaches. Ravin Yael and Leacock Claudia (ed.). Oxford University Press. New York. 178-204.

[23] **Earl, Lois L. (1973).** Use of word government in resolving syntactic and semantic ambiguities. Information Storage and Retrieval, 9, 639-664.

[24] **Fillmore, Charles J. y Atkins, Beryl T. S. (1991)**. Invited lecture presented at the 29th Annual Meeting of the Association for Computational Linguistics, 18-21 June 1991, Berkeley, California.

[25] **Firth, J. R. (1957).** Modes of meaning. Papers in Linguistics 1934-51. Oxford University Press, Oxford, United Kingdom. 190-215.

[26] **Fox, B. A. (1987).** Discourse structure and anaphora: written and conversational English. Cambridge University Press, Cambridge. USA.

[27] **Gale, William A.; Church, Kenneth W. y Yarowsky, David (1992).** One sense per discourse. Proceedings of the Speech and Natural Language Workshop, San Francisco, Morgan Kaufmann, 233-37.

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

[29] **Galicia-Haro Sofía N., Bolshakov I. A. y Gelbukh A. F. (1999).** Un modelo de descripción de la estructura de las valencias de verbos españoles para el análisis automático de textos.

[30] **Gelbukh, Alexander (1997).** Using a semantic network for lexical and syntactical disambiguation. CIC-97, nuevas aplicaciones e Innovaciones Tecnológicas en Computación, Simposio Internacional de Computación, Mexico City, Mexico, pp. 352-366.

[31] **Gelbukh, Alexander (2000).** Computational Processing of Natural Language: Tasks, Problems and Solutions. Congreso Internacional de Computación en México D.F., Nov 15-17, 2000.

[32] **Gelbukh, Alexander y Sidorov Grigori (1999a).** On Indirect Anaphora Resolution. In: Proceedings of PACLING-99, Waterloo, Ontario, Canada, 181-190.

[33] **Gelbukh, Alexander y Sidorov Grigori (1999b)**. A Thesaurus-based Method for Indirect Anaphora Resolution. Revised version of On Indirect Anaphora Resolution In: Proceedings of PACLING-99.

[34] **Gelbukh, Alexander y Sidorov Grigori (2001).** La estructura de dependencias entre las palabras en un diccionario explicativo del español: resultados preliminares

[35] **Genesereth, M. R (1997).** Keller, A. M. & Duschka, O. Infomaster: An Information Integration System. In Proceedings of 1997 ACM SIGMOD Conference, May 1997.

[36] **Ghazfan (1996).** Toward meaningful Bayesian networks for information retrieval systems. In Proceedings of the IPMU´96 Conference, pages 841-846.

[37] **Global Reach (2002)**, http://global-reach.biz.

[38] **Guzmán Arenas, Adolfo (1999b)**. Finding the main themes in a spanish document. Journal Expert Systems with Applications, Vol 14, Nº 1/2. January/February, 139-148.

[39] **Haas, W. (1966).** Linguistic relevance. In Bazell, C.E. et al. (Eds.), In Memory of J.R. Firth, Longman, London, 116-48.

[40] **Hale (1997).** Michael L. Mc. A comparison of WordNet and roget's taxonomy for measuring semantic similarity.

[41] **Halliday, M.A.K. (1961)** Categories of the theory of grammar. Word, 17, 241-92.

[42] **Halliday, M.A.K. (1966)** Lexis as a linguistic level. In Bazell, C.E. et al. (Eds.), In Memory of J.R. Firth, Longman, London, 148-63.

[43] **Harris, Zellig S. (1954).** Distributional Structure. Word, 10, 146-162.

[44] **Hayes, Philip J. (1977a).** On semantic nets, frames and associations. Proceedings of the 5th International. Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 99-107.

[45] **Hayes, Philip J. (1977b)**. Some association-based techniques for lexical disambiguation by machine. Doctoral dissertation, Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne.

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

[47] **Hindle, Donald y Rooth, Mats (1993).** Structural Ambiguity and Lexical Relations. Computational Linguistics, 19(1), 103-120.

[48] **Hirst, Graeme (1981).** Anaphora in Natural Language Understanding. Springer Verlag, Berlin.

[49] **Hirst, Graeme (1987).** Semantic interpretation and the resolution of ambiguity. Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom, 263pp.

[50] **Hirst, Graeme (1998)** Chapter 13. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. WordNet An electronic Lexical Database. Edited by Christiane Fellbaum. The MIT Press. Cambridge, Massachusetts, London, England.

[51] **Hjemslev, Louis (1953).** Prolegomena to a theory of language. Translated from Danish. Indiana University, Bloomington, Indiana.

[52] **HLTTeam (2002).** Language Engineering. Harnessing the Power of Language. http://www.hltcentral.org/usr_docs/Harness/harness-en.htm.

[53] **Huang Yang (2000).** Anaphor: A Croos-linguistic Approach. Oxford University Press, New York, USA.

[54] **Jennings, N.R. and M.Wooldridge, (1998).** Applications of Intelligent Agents. Agent Technology. Foundations, Applications, and Markets. Eds.: N.R.Jennings and M.Wooldridge. Springer, 1998.

[55] **Jensen, Karen y Binot, Jean-Louis (1987).** Disambiguating prepositional phrase attachments by using on-line dictionary definitions. Computational Linguistics, 13(3/4), 251-260.

[56] **Jorgensen, Julia (1990).** The psychological reality of word senses. Journal of Psycholinguistic Research, 19, 167-190.

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

[58] **Kilgarriff, Adam (1998)**. I don't believe in word senses. In Computers and the Humanities..

[59] **Krovetz, Robert y Croft, William Bruce (1992),** Lexical Ambiguity and Information Retrieval. ACM Transactions on Information Systems, 10(2), 115-141.

[60] **Lawrence Steve (2000),** El Acceso a la Información en la Web Limitado y Desigual. NEC Research Institute, http://www.neci.nec.com/. XE "Enl

[61] **Leacock, Claudia; Miller, George A. y Chodorow, Martin (1998).** Using corpus statistics and WordNet relations for sense identification.

[62] **Letch, Charley. (1992).** Información Tsunami: Un futurista mira en retrospectiva, Primera Edición, Editorial Limusa, Colección Megabyte, México D.F.

[63] **Lesk, Michael (1986).** Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June 1986, 24-26.

[64] **Lewis, David D. (1995).** Evaluating and Optimizing Autonomous Text Classification Systems. Proceeding of the 18 Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. 1995.

[65] **Lewis, David D. (1995a).** Active by Accident: Relevance Feedback in Information Retrieval. From the unpublished working notes of the 1995 AAAI Fall Symposium on Active Learning. 1995.

[66] **Litkowski, Kenneth C. (1997).** Desiderata for tagging with WordNet sysnsets or MCAA categories. ACL-SIGLEX Workshop “Tagging Text with Lexical Semantics: Why, What, and How?” April 4-5, 1997, Washington, D.C., USA, 12-17.

[67] **Lyons, John (1966).** Firth's theory of meaning. In Bazell, C.E. et al. (Eds.). In Memory of J.R. Firth, Longman, London, 288-302.

[68] **Lyons, John (1977).** Semantics. Cambridge University Press, Cambridge, England.

[69] **Malakhovski, L. V. (1987).** Homonyms in English dictionaries. In Burchfield, R. W. (Ed.), Studies in Lexicography. Oxford University Press, Oxford, United Kingdom, 36-51.

[70] **Manning Christopher D. y Schütze, Hinrich (1999).** Foundations of Statistical Natural Language Processing. MIT Press, London, England. (2th printing 2000).

[71] **McIntosh, A. (1966).** Patterns and ranges. Papers in General, Descriptive, and Applied Linguistics. Longman, London, 183-99.

[72] **McRoy, Susan W. (1992).** Using multiple knowledge sources for word sense discrimination. Computational Linguistics, 18(1)1-30.

[73] **Meillet, Antoine (1926).** Linguistique historique et linguistique générale. Vol. 1. Champion, Paris, 351pp. (2nd édition).

[74] **Mihalcea R. and Moldovan D. (1999).** A Method for word sense disambiguation of unrestricted text. Proc 37th Annual Meeting of the ACL 152-158, Maryland, USA.

[75] **Mitkov, R. (1998).** Evaluating anaphora resolution approaches. In: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2). Lancaster, UK.

[76] **Mitkov, R. (2001).** Oustanding Issues in Anaphora Resolution. In: Proceedings of Second International Conference, CICLing 2001, Mexico City, México, 18-24 February. Alexander Gelbukh (Ed.). Lecturer Notes in Computer Science LNCS 2004 Springer 110-125.

[77] **Montoyo, Andrés (2001).** Método badaso en Marcas de Especificidad para WSD, Grupo de Procesamiento del Lenguaje y Sistemas de Información. Universidad de Alicante, España.

[78] **Nida, E. (1966).** A review of Martinet, André (1950). Morphology : The descriptive analysis of words. Word, 6(1), 84-87.

[79] **Pereira, Fernando y Tishby, Neftali (1992).** Distributional similarity, phase transitions and hierarchical clustering. Working notes of the AAAI Symposium on Probabilistic Approaches to Natural Language, October 1992, Cambridge, Massachusetts, 108-112.

[80] **Pereira, Fernando; Tishby, Neftali; y Lee, Lilian (1993).** Distributional clustering of English. Proceeedings of the 31st Annual Meeting of the Association for Computational Linguistics, June 1993, 22-26. Ohio State University, Colombus, Ohio, 183-190.

[81] **Pimienta, Daniel (2000).** Representación de las lenguas y culturas latinas en la Internet, Fundación Redes y Desarrollo. Encuentro Sociedad y Tecnología, Santiago de Chile, 14-15/12/2000, http://funredes.org.

[82] **Quine, Willard V. (1960).** Word and object, The MIT Press, Cambridge, Massachusetts.

[83] **Ravin Yael y Leacock Claudia (2000)**. Polysemy: an overview. In Polysemy: Theoretical and Computational Approaches. Ravin Yael and Leacock Claudia (ed.). Oxford University Press. New York. 1-29.

[84] **Resnik P. (1995).** Disambiguating noun groupings with respect to WordNet senses. Proc. Third Workshop on Very Large Corpora. 54-68. Cambridge, MA.

[85] **Resnik P. (1999).** Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. In Journal of Artificial Intelligence Research 11. 95-130.

[86] **Ribeiro (1996).** A belief network model for IR. In Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. SIGIR´96, August 18-22, 1996, Zurich, pages 253-260. ACM.

[87] **Rigau g., Atserias J. and Aguirre E. (1997).** Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. Proc 35th annual Meeting of the ACL, 48-55, Madrid, Spain.

[88] **Robins, R. H. (1987).** Polysemy and the lexicographer. In Burchfield, R. W. (Ed.), Studies in Lexicography. Oxford University Press, Oxford, United Kingdom, 52-75.

[89] **Salton, Gerard (1968).** Automatic Information organization and Retrieval. McGraw-Hill, New Cork.

[90] **Salton, Gerard y McGill, M. (1983).** Introduction to Modern Information Retrieval. McGraw-Hill, New York.

[91] **Saracevic, Tefko (1975).** Relevance: a review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6):321-343, 1975. Also reprinted in [SJW97], pp. 143-165.

[92] **Saracevic, T. (1995).** A taxonomy of values for library and information services. Rutgers University, New Brunswick.

[93] **Schank, Roger C. y Abelson, Robert P. (1977).** Scripts, Plans, Goals and Understanding. Lawrence Erlbaum, Hillsdale, New Jersey.

[94] **Schütze, Hinrich y Pedersen, Jan (1995).** Information retrieval based on word senses. Proceedings of SDAIR'95. April 1995, Las Vegas, Nevada.

[95] **Schütze, Hinrich (2000).** Disambiguation and Connectionism. In Polysemy: Theoretical and Computational Approaches. Ravin Yael and Leacock Claudia (ed.). Oxford University Press. New York. 205- 219.

[96] **Seneff, Stephanie (1992).** TINA, A natural language system for spoken language applications. Computational Linguistics, 18(1), 61-86.

[97] **Shoham, Y, (1993).** Agent-Oriented Programming. Artificial Intelligence, 60(1), pp: 51-92.

[98] **Sidorov Grigori y Gelbukh, Alexander (1999).** Demostrative Pronouns as Markers of Indirect Anaphora.

[99] **Simpson, Greg B. y Burgess, Curt (1988).** Implications of lexical ambiguity resolution for word recognition and comprehension. In Small, Steven; Cottrell, Garrison W.; and Tanenhaus, Michael K. (Eds.) (1988). Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence. Morgan Kaufman, San Mateo, California, 271-288.

[100] **Sinclair, John (1966).** Beginning the study of lexis. In Bazell, C.E. et al. (Eds.). In Memory of J.R. Firth. Longman, London, 410-31.

[101] **Sinclair, John (Ed.) (1987).** Looking up: An account of the COBUILD project in lexical computing. Collins, London, 182pp.

[102] **Sproat, Richard; Hirschberg, Julia; y Yarowsky, David (1992).** A corpus-based synthesizer. Proceedings of the International Conference on Spoken Language Processing, Banff. Alberta, Canada, October 1992.

[103] **Stetina J., Kurohashi S. and Nagao M. (1998).** General word sense disambiguation method based on full sentencial context. In Usage of WordNet in Natural Language Processing. COLING-ACL Workshop, Montreal, Canada.

[104] **Stock, Penelope F. (1983).** Polysemy. Proceedings of the Exeter Lexicography Conference, 131-140.

[105] **Stone, Philip J. (1969).** Improved quality of content analysis categories: Computerized-disambiguation rules for high-frequency English words. In Gerbner, George; Holsti, Ole, Krippendorf, Klaus; Paisley, William J.; and Stone, Philip J. (Eds.), The Analysis of Communication Content, John Wiley and Sons, New York, 199-221.

[106] **Sussna M. (1993).** Word sense disambiguation for free-text indexing using a massive semantic network. Proc. Second International CIKM, 67-74, Airlington, A.

[107] **Towell, Geoffrey y Voorhees, Ellen (1998).** Disambiguating highly ambiguous words.

[108] **Turtle and Croft (1990).** Inference networks for document retrieval. In SIGIR´90, 13th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5-7 September 1990, Proceedings, pages 1-24. ACM, 1990.

[109] **Van Buren, P. (1967).** Preliminary aspects of mechanisation in lexis. CahLex, 11, 89-112; 12, 71-84.

[110] **Voorhees, Ellen M.; Claudia Leacock, y Geoffrey Towell (1995).** Learning context to disambiguate word senses. In Thomas Petsche; Stephen José Hanson; and Jude Shavlik, eds., Computational Learning Theory and Natural Learning Systems. MIT Press, Cambridge, Massachusetts.

[111] **Voorhees, Ellen M. (1993).** Using WordNet to disambiguate word senses for text retrieval. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 27 June-1 July 1993, Pittsburgh, Pennsylvania, 171-180.

[112] **Weaver, Warren (1949).** Translation. Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, William N. y Booth, A. Donald (1955) (Eds.), Machine translation of languages. John Wiley and Sons, New York, 15-23.

[113] **Weinreich, Uriel (1980).** On semantics. University of Pennsylvania Press, 128pp.

[114] **Whittemore, Greg; Ferrara, Kathleen; y Brunner, Hans (1990).** Empirical studies of predictive powers of simple attachment schemes for post-modifier prepositional phrases. Proceedings of the 28th Annual Meeting of Association for Computational Linguistics, 6-9 June 1990, Pittsburgh, Pennsylvania, 23-30.

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

[116] **Wilks, Yorick A. (1975).** Preference semantics. In Keenan, E. L. III (Ed.), Formal Semantics of Natural Language. Cambridge University Press, 329-348.d.

[117] **Wiks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. (1993).** Providing Machine Tractable dictionary tools. In Semantics and the lexicon (Pustejowsky J. Ed.) 341-401.

[118] **Wilks, Yorick A. (1998).** Senses and texts. In Computers and the Humanities

[119] **Wilks, Yorick A. y Stevenson, Mark (1996).** The grammar of sense: Is word sense tagging much more than part- of-speech tagging? Technical Report CS-96-05, University of Sheffield, Sheffield, United Kingdom.

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

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

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

[123] **Yarowsky, David (1997).** Homograph disambiguation in text-to-speech synthesis. In Van Santen, Jan T. H.; Sproat, Richard; Olive, Joseph P.; and Hirschberg, Julia. Progress in Speech Synthesis. Springer-Verlag, New York, 157-172.

[124] **Yngve, Victor H. (1955).** Syntax and the problem of multiple meaning. In Locke, William N. and Booth, A. Donald (Eds.), Machine translation of languages. John Wiley & Sons, New York, 208-226.

Índice de términos

Índice de términos

En este índice se presenta la ubicación dentro del documento de algunos términos y temáticas ordenadas alfabéticamente para su rápida localización.

Índice de términos

Anexos

En este apartado se presenta información adicional que puede ayudar a comprender las ideas expresadas.

Anexos

Anexo 1. Interfaz

El propósito del algoritmo de normalización morfológica es encontrar su lema, para una forma de palabra dada (forma normalizada).

Se tienen formas de palabras en una tabla donde para cada forma se guarda su lema e información gramática. Hay métodos más sofisticados pero éste es el más simple. Por ahora el diccionario morfológico contiene alrededor de 500,000 formas de palabras.

Ejemplo del diccionario morfológico y su estructura se encuentran en Ilustraciones 1 y 2.

En este nivel no se resuelve la homonimia, por ejemplo, para la forma *trabajo* se generan dos lemas: *trabajar* y *trabajo.*

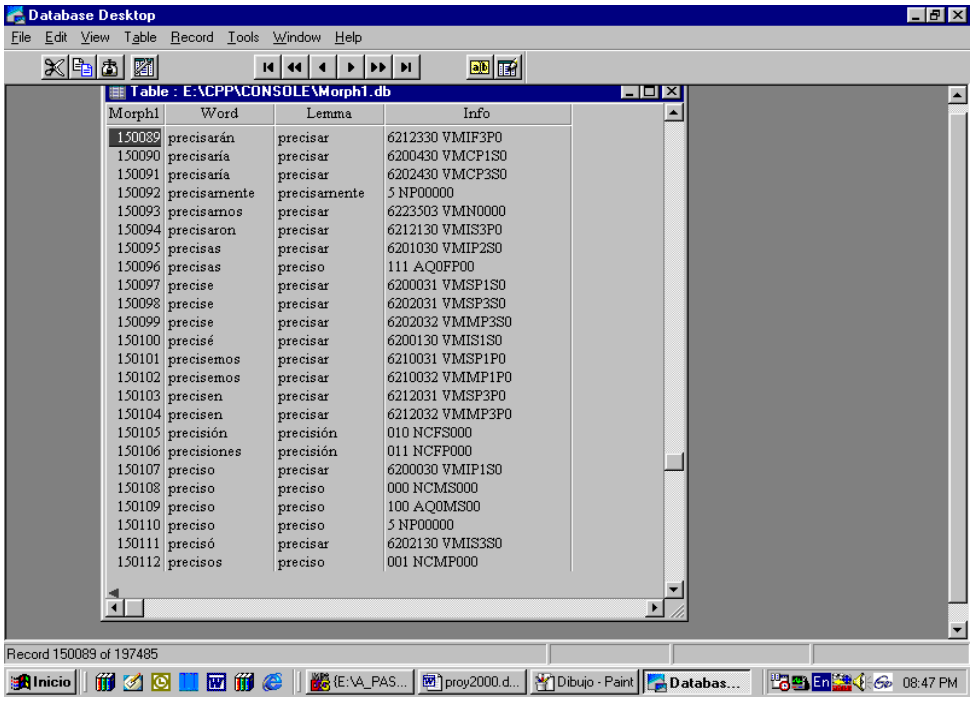


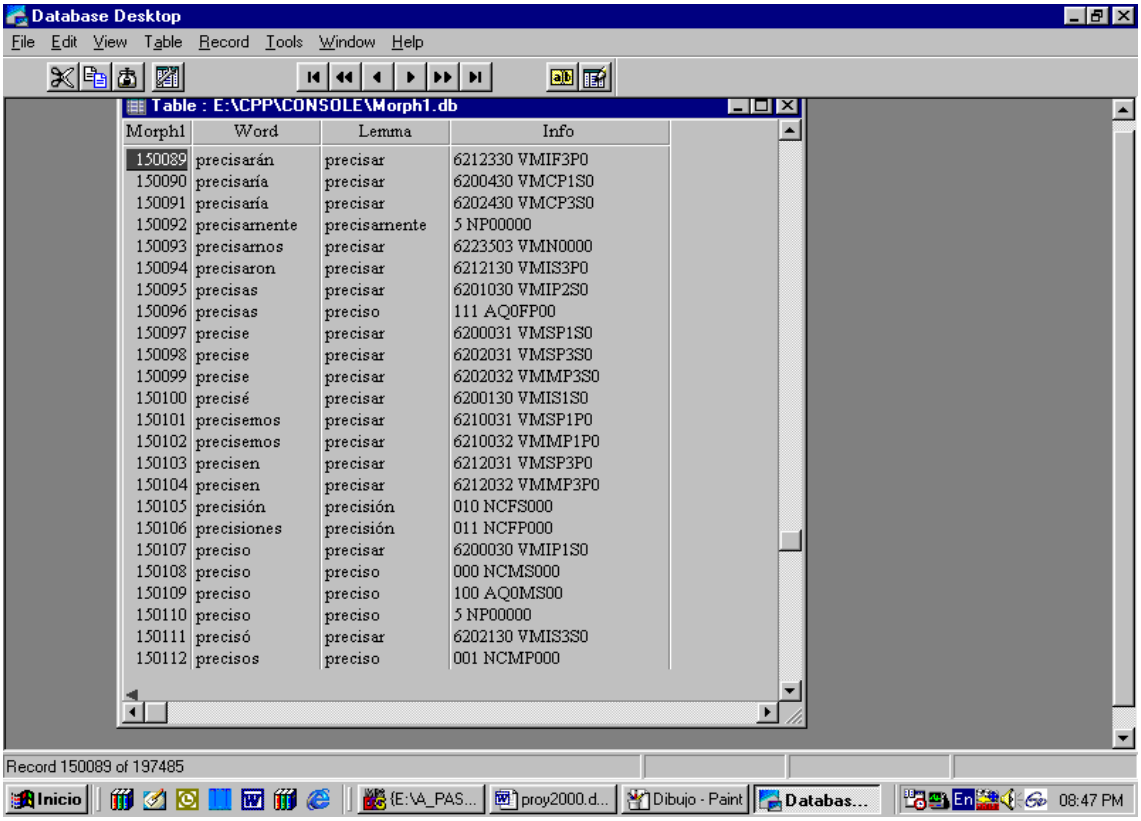
Figura 10. Diccionario morfológico para español

Se aplica este módulo al diccionario explicativo de español. Cambiando todas las formas de palabras por sus lemas (posiblemente más de una) y se guardan los resultados en la misma base de datos.

De hecho este algoritmo funciona junto con el algoritmo de desambiguación morfológica para no guardar los datos innecesarios.

Anexo 2. Algoritmo de desambiguación morfológica

El propósito del algoritmo de desambiguación morfológica es resolver la homonimía morfológica. Para cumplir esta tarea se usan la heurística sintáctica.



Morph1	Word	Lemma	Info
150089	precisarán	precisar	6212330 VMIF3P0
150090	precisaría	precisar	6200430 VMCP1S0
150091	precisaría	precisar	6202430 VMCP3S0
150092	precisamente	precisamente	5 NP00000
150093	precisamos	precisar	6223503 VMN0000
150094	precisaron	precisar	6212130 VMIS3P0
150095	precisas	precisar	6201030 VMIP2S0
150096	precisas	preciso	111 AQ0FP00
150097	precise	precisar	6200031 VMSP1S0
150098	precise	precisar	6202031 VMSP3S0
150099	precise	precisar	6202032 VMMP3S0
150100	precisé	precisar	6200130 VMIS1S0
150101	precisemos	precisar	6210031 VMSP1P0
150102	precisemos	precisar	6210032 VMMP1P0
150103	precisen	precisar	6212031 VMSP3P0
150104	precisen	precisar	6212032 VMMP3P0
150105	precisión	precisión	010 NCFS000
150106	precisiones	precisión	011 NCFP000
150107	preciso	precisar	6200030 VMIP1S0
150108	preciso	preciso	000 NCMS000
150109	preciso	preciso	100 AQ0MS00
150110	preciso	preciso	5 NP00000
150111	precisó	precisar	6202130 VMIS3S0
150112	precisos	preciso	001 NCMP000

Figura 11. Estructura del diccionario morfológico para español

Se usan las siguientes heurísticas sintácticas:

- La primera heurística elimina la hipótesis de sustantivo que tiene un verbo homónimo si el artículo que se encuentra antes del sustantivo no tiene concordancia con el sustantivo, por ejemplo, *el trabajo*, *trabajo* puede ser verbo ó sustantivo, pero si, digamos, el artículo tenía la forma femenina (*la*), entonces la hipótesis de sustantivo sería eliminada.

- La segunda heurística elimina la hipótesis de sustantivo que tiene un verbo homónimo si antes tenemos las clíticas verbales. Por ejemplo, *me llamas.* En este caso, *llamas* podría ser el sustantivo *llama* o el verbo *llamar*. Presencia de la clítica indica que es verbo.
- La tercera heurística elimina la hipótesis de verbo si antes tenemos los artículos "*un* / *una*". Esta situación es imposible para el verbo, por ejemplo, *un trabajo*. Aquí *trabajo* es el sustantivo.
- La siguiente heurística usa el hecho de que la preposición no puede encontrarse inmediatamente antes del verbo. Por lo tanto podemos eliminar la hipótesis del verbo si la palabra a la izquierda es la preposición. Por ejemplo, *de trabajo.*
- La otra heurística trata de homonimia entre el sustantivo y el adjetivo, por ejemplo *transmisión local, local* podría ser el adjetivo o el sustantivo. En caso de que la palabra anterior sea el sustantivo que no tiene el homónimo adjetivo, podemos eliminar la hipótesis de sustantivo. En este caso *local* será solo el adjetivo.
- La siguiente heurística debe procesar las situaciones de tiempos verbales compuestos, por ejemplo, "*había visto*", donde *había* puede ser la forma de verbo "*haber*" o solo un elemento gramático. La heurística simplemente verifica que después de "*haber*" esté el participio pasado y este caso ignora el verbo.
- La otra heurística trata de resolver la situación de dos sustantivos, uno de los cuales es *Pluralia tantum* y el otro es normal, por ejemplo, *botón-botones*. En este caso hay que evitar la aparición de la lema en forma de singular.
- La penúltima heurística se distingue de las demás por ser la heurística estadística. De los homónimos se deja el homónimo que tiene más frecuencia en el diccionario. El argumento a favor es que la homonimía se encuentra sólo en algunas formas de las palabras, entonces la palabra que aparece en el texto sólo gracias a homonimía debe tener menor frecuencia.
- La última heurística se basa en la específica del texto. Porque es en el diccionario donde deben aparecer más frecuentemente las formas de sustantivos que de adjetivos, de adjetivos más frecuente que de verbos, etc. Por lo tanto, como último paso se puede resolver la homonimía usando este criterio.