RESUMEN DE TESIS DOCTORAL

Recuperación de Información con Resolución de Ambigüedad de Sentidos de Palabras para el Español

Information Retrieval with Word Sense Disambiguation for Spanish

Graduated: Yoel Ledo Mezquita
Centro de Investigación en Computación (CIC-IPN)
Av. Juan de Dios Bátiz sn esq. Miguel Othón de Mendizábal C. P. 07738 México D. F.,
yledo@yahoo.com
Graduated on June 23, 2006

Advisor: Grigori Sidorov
Centro de Investigación en Computación (CIC-IPN)
Av. Juan de Dios Bátiz sn esq. Miguel Othón de Mendizábal C. P. 07738 México D. F.,
www.cic.ipn.mx/~sidorov

Advisor: Alexander Gelbukh
Centro de Investigación en Computación (CIC-IPN)
Av. Juan de Dios Bátiz sn esq. Miguel Othón de Mendizábal C. P. 07738 México D. F.,
www.cic.ipn.mx/~sidorov

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 son debidas a los distintos sentidos que tienen las palabras, lo cual se le 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, en función de sus posibles sentidos semánticos. Las aportaciones de este artículo consisten en el desarrollo de un nuevo método de desambiguación de sentidos de palabras usando grandes recursos léxicos (diccionarios explicativos, diccionarios de sinónimos, WordNet).
Palabras clave: *sentidos de palabras, contexto, diccionarios, algoritmo de Lesk.*

Abstract
One of the problems of information retrieval in Internet and digital libraries is low precision: a high number of retrieved documents of low relevance. For example, a person looks for information about *jaguars* (the animal) and the documents retrieved are about the *model of a car*. This problem arises due to ambiguity of different senses of words. The task of determining the correct interpretation of a word in its context is known as Word Sense Disambiguation (WSD) task. It employs a linguistic mechanism that detects the most suitable sense of a word, according to the context where the word is used, choosing of its possible senses. In this paper, a new method for word senses disambiguation is proposed based on additional linguistic information for the words in the context available from the large lexical resources, like explanatory dictinary, synonym dictionary, WordNet.
Keywords: *word senses, context, dictionaries, Lesk algorithm.*

1 Introducción

1.1 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 a 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 que contribuyan a la solución de esta problemática, no ha sido muy investigado para la recuperación de información y en la lingüística computacional. Producir conocimiento sobre las búsquedas inteligentes para los servicios en línea, las condiciones que la determinan y sus mecanismos de procesamiento apoyado 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 nos 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 su 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.

En el presente trabajo nos centramos 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 son debidas a los distintos sentidos que tienen las palabras.

1.2 Descripción del problema

La recuperación de información 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 son expresadas 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 documentos y las necesidades de información, son dependientes de las preferencias e interpretaciones subjetivas del usuario. Además, esta relación es inherentemente no formalizable [Saracevic, 1995].

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

Uno de los problemas en los portales de recuperación de información en Internet (por ejemplo, los portales dinámicos Altavista, Google, Yahoo, etc.) y en bibliotecas digitales (por ejemplo, Biblioteca del Congreso de los EEUU, etc.) 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 más importantes problemas de investigación en el procesamiento del lenguaje natural [Wilks and Stevenson, 1996]. Es esencial para las aplicaciones que requieren la comprensión del lenguaje y de mensajes, comunicación hombre-máquina, la recuperación de información y otros. Es requerido 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 quiere 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 geenración de 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 limitarse nada más a sacar las primeras líneas de los párrafos.
- Reconocimiento de voz. Esta es una de las aplicaciones del 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: para identificar al usuario o para procesar lo que el usuario dicte. Y existen ya programas comerciales accesibles por los usuarios por ejemplo: Dragon Naturally Speaking.
- Recuperación de información. Un claro ejemplo de esta aplicación sería el siguiente: una persona llega a la computadora y le dice en lenguaje natural (oral o escrito) qué es lo que busca; ésta busca y le dice qué es lo que tiene referente al tema.

En los últimos diez años se han multiplicado las investigaciones para desambiguar palabras automáticamente y crear métodos para identificar los problemas y aplicar las soluciones encontradas. Sin embargo, 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 los sentidos de las palabras usando grandes recursos léxicos para ser aplicado en la recuperación de la información y en la navegación en hipertexto.

1.3 Objetivo general

El disponer de servicios de recuperación inteligente de información eficientes permitiría mejorar la respuesta a los usuarios que buscan información. En base a esta suposición, el objetivo general de esta investigación es diseñar un nuevo método de desambiguación de los sentidos de las palabras usando grandes recursos léxicos que mejore la pertinencia de la información recuperada.

2 Antecedentes

En cierto sentido el trabajo de WSD ha vuelto recientemente a los métodos empíricos y a los análisis basados en corpus que caracterizan algunos de los esfuerzos iniciales por resolver el problema. Con mayores recursos y métodos estadísticos reforzados a su disposición, los investigadores están mejorando los resultados de los pioneros, pero parece que están acercaándose 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 de determinar o incluso definir el sentido de la palabra, y esto no parece ser 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, dos son los que más influencia han tenido en el área:

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

2.1 Enfoque basado en métodos estadísticos

En este enfoque no se usan algunas fuentes adicionales de conocimiento para la resolución de la ambigüedad [Manning and Shutze, 1999], lo 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, por lo general 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 tenemos: 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 más relevantes de información basados en las redes bayecianas .

El primero, denominado Modelo de las redes de interferencias (*Inference Network Model*), 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 los 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 y, contiene nodos consulta y nodos término (los arcos van de los nodos término a los nodo consulta). [Turtle and 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. Una vez que todas las propagaciones hayan finalizado, se genera el correspondiente ordenamiento de documentos.

Estrechamente relacionado con este trabajo, está el conocido como método 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, los documentos se ordenan según las probabilidades de pertenecer a la respuesta. Para ello, se instancian los nodos de la consulta, propagando sólo 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 Modelo de las redes de creencia (*Belief Network Model*), donde se consideran únicamente dos tipos de nodos, documentos y términos, enlazándolos 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 sus probabilidades.

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].

2.2 Enfoque basado en las fuentes adicionales de conocimiento

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

Durante los últimos años se han realizado algunas investigaciones sobre WSD basada 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 las palabras del contexto y sus definiciones del diccionario explicativo.

Cowie [Cowie *et al*., 1992] describe un método para resolver la ambigüedad léxica de textos basado en la definición dada en *Longman's Dictionary of Contemporary English* (Diccionario de ingles contemporáneo de Longman) 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 consistía 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 similaridad semántica para las palabras en la jerarquía WordNet. En [Aguirre and Rigau, 1996] se combinan un conjunto de algoritmos no supervisados para desambiguar nombres del corpus Semcor. En [Rigau *et a*l., 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 *et al*., 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 and 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.

Al valorar los dos enfoques analizados, el enfoque mediante métodos estadísticos y el enfoque mediante fuentes de conocimiento, decidimos 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.

3 Método Propuesto

La idea del algoritmo de Lesk original [Lesk, 1986] se basa en la búsqueda de intersección de las definiciones de un diccionario explicativo con las definiciones de palabras del contexto en el mismo diccionario. La idea del algoritmo de Lesk simplificado consiste en la búsqueda de las intersecciones de definición de la palabra en cuestión con las palabras del contexto.

Estamos proponiendo un método de desambiguación de los sentidos de las palabras que es una combinación de ambos métodos y utiliza adicionalmente la información léxica obtenida de diferentes diccionarios. Además usamos la posición relativa de las palabras del contexto para ponderar sus pesos.

El algoritmo realiza varios pasos.

Primero, hacemos la normalización morfológica de las palabras del documento resolviendo la homonimía morfológica usando un conjunto de las heurísticas sintácticas. Cada palabra se sustituye por su forma normalizada (que se llama lema), por ejemplo, *trabajo, trabaja, trabajé, trabajaron* se sustituye por el infinitivo *trabajar*.

Después, se eliminan las palabras auxiliares del documento; tales como preposiciones, artículos, verbos auxiliares, pronombres. En este caso, después de este procedimiento el documento contiene solamente verbos, adjetivos, sustantivos y adverbios.

Realizamos desambiguación de los sentidos de las palabras en un dominio local limitado a cierto número de palabras.

Hicimos los experimentos con diferentes tamaños de la ventana. Nuestros experimentos demostraron que los mejores resultados se obtienen con la ventana de tamaño menor de 17 palabras, véase **Ilustración 2**. Entonces, usamos este tamaño de la ventana. El método pondera la influencia de la palabra de contexto dependiendo de la distancia de la palabra en cuestión: Más cercana es la palabra, mayor influencia tiene. De manera empírica llegamos a la conclusión que los coeficientes exponenciales son mejores para ponderar dicha influencia dentro de la ventana mencionada.

En el siguiente paso del algoritmo, para cada palabra con varios sentidos se calculan valores utilizando ambos algoritmos basados en el método de Lesk. Después aplicamos el mismo procedimiento para los sinónimos obtenidos de un diccionario de sinónimos y para las palabras que forman otras relaciones léxicas (hiperónimos, hipónimos, etc.) obtenidas de WordNet. Cada valor se pondera según la distancia de la palabra en cuestión. Sumamos los valores obtenidos para cada sentido aplicando diferentes pesos a los valores retornados: para los algoritmos de Lesk utilizamos el peso 1, para los sinónimos 0.9, y para otras relaciones léxicas 0.8. Los valores óptimos de estos pesos se obtuvieron haciendo las series de experimentos.

4 Resultados Experimentales

En total hicimos experimentos para 4,287 sentidos de 872 palabras. Usamos 80 contextos que contenían en total 1,293 palabras. Hicimos tres tipos de evaluaciones usando el algoritmo de Lesk original, el algoritmo de Lesk simplificado, y el método propuesto.

Ejemplos de los datos de entrada se presentan en Tabla 1.

Tabla 1. Ejemplos de datos de entrada

No.	Contexto	Palabras para desambiguar	Sentidos 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*	5	14
3.	*No se pierda las razones para suscribirse al mejor diario digital.*	6	50
4.	*Participa en nuestro concurso de fotografía digital*	4	12

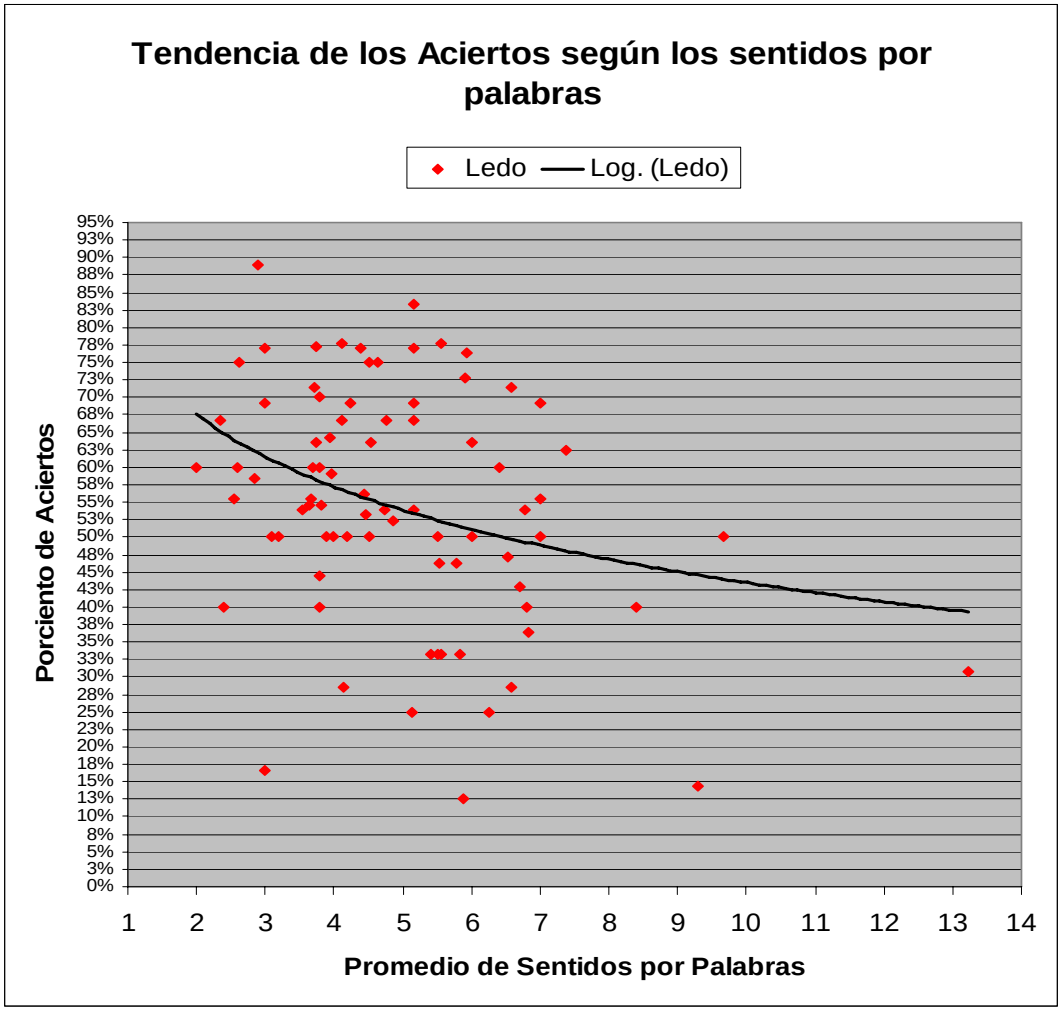

Ilustración 1. Tendencias de aciertos según los sentidos por las palabras.

En la Ilustración 1 se presentan las tendencias de aciertos según el número de sentidos de palabras en los experimentos realizados. Obviamente, con mayor número de sentidos es más dificil encontrar un sentido correcto, sobretodo que los sentidos empiezan a reflejar las diferencias muy finas.

Ahora, como un ejemplo, vamos a presentar los resultados experimentales para el contexto "*Dejar a un gato normalmente sociable en una pensión o en la clínica*". Los resultados obtenidos usando métodos diferentes se presentan en Tablas 2 a 7.

Tabla 2. Contribución de palabras del contexto usando el método de Lesk original (palabra "*sociable*")

Palabra 3	**sociable**	Sentidos: 1
		0
dejar	14	0
gato	7	0.021
normalmente	0	0
pensión	4	0.032
clínica	5	0
Totales		0.053

Tabla 3. Contribución de palabras del contexto usando el método de Lesk original (palabra "*pensión*")

Palabra 4	**pensión**	Sentidos: 4			
		0	1	2	3
dejar	14	0.000	0.032	0.000	0.000
gato	7	0.019	0.019	0.000	0.000
normalmente	0	0.000	0.000	0.000	0.000
pensión	1	0.000	0.141	0.000	0.000
clínica	5	0.000	0.000	0.000	0.000
Totales		**0.019**	**0.192**	**0.000**	**0.000**

Tabla 4. Contribución de palabras del contexto usando el método de Lesk simplificado (palabra "*sociable*")

Palabra 3	**sociable**	Sentidos: 1
		0
dejar	14	0.000
gato	7	0.000
normalmente	0	0.000
pensión	4	0.000
clínica	5	0.000
Totales		0.000

Tabla 5. Contribución de palabras del contexto usando el método de Lesk simplificado (palabra "*pensión*")

Palabra 4	**pensión**	Sentidos: 4			
		0	1	2	3
dejar	14	0.000	0.000	0.000	0.000
gato	7	0.019	0.000	0.000	0.000
normalmente	0	0.000	0.000	0.000	0.000
pensión	1	0.000	0.000	0.000	0.000
clínica	5	0.000	0.000	0.000	0.000
Totales		**0.000**	**0.000**	**0.000**	**0.000**

Table 6. Contribución de palabras del contexto usando el método propuesto (palabra "s*ociable*")

Palabra 3	**sociable**	Sentidos: 1
		0
dejar	14	0.038
gato	7	0.043
normalmente	0	0.000
pensión	4	0.032
clínica	5	0.079
Totales		**0.192**

Table 7. Contribución de palabras del contexto usando el método propuesto (palabra "*pensión*")

Palabra 4	**pensión**	Sentidos: 4			
		0	1	2	3
dejar	14	0.090	0.135	0.065	0.065
gato	7	0.283	0.059	0.039	0.098
normalmente	0	0.072	0.000	0.000	0.000
pensión	1	0.152	0.141	0.000	0.000
clínica	5	0.152	0.082	0.000	0.082
Totales		**0.749**	**0.412**	**0.104**	**0.245**

Nótese que los resultados obtenidos con el método propuesto son mejores que en caso de los otros métodos. En la siguiente tabla se presentan los resultados globales.

Tabla 8. Comparación del método propuesto con los métodos de base

Contexto			**Lesk Original**		**Lesk Simplificado**		**Método propuesto**	
No	**W**	**Sentidos**	S	P (%)	S	P (%)	S	P (%)
1.	6	31	2	33	2	33	5	83
2.	5	14	3	50	2	33	4	67
3.	6	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

En la tabla 8, *W* significa *palabras (word)*, *S* significa *éxito (success)*, y *P* significa p*recisión*.

El método propuesto obtuvo 63% de precisión, que es 13% mejor que el método original de Lesk que obtuvo 50% y 22% mejor que el método de Lesk simplificado que obtuvo 41%.

Conclusiones

Nuestro método realiza 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 varios recursos léxicos.

Los resultados obtenidos demuestran que nuestro método obtuvo mejor precisión que los otros dos métodos anteriores.

Nuesto método propone una nueva forma de determinar la semejanza de contextos 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.

Nuestro método hace 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.

La disponibilidad de un método de desambiguación de sentidos de palabras para la recuperación inteligente de información en buscadores de tanta necesidad en la actualidad por el gran volumen de información que existe y se consumen diariamente por las personas

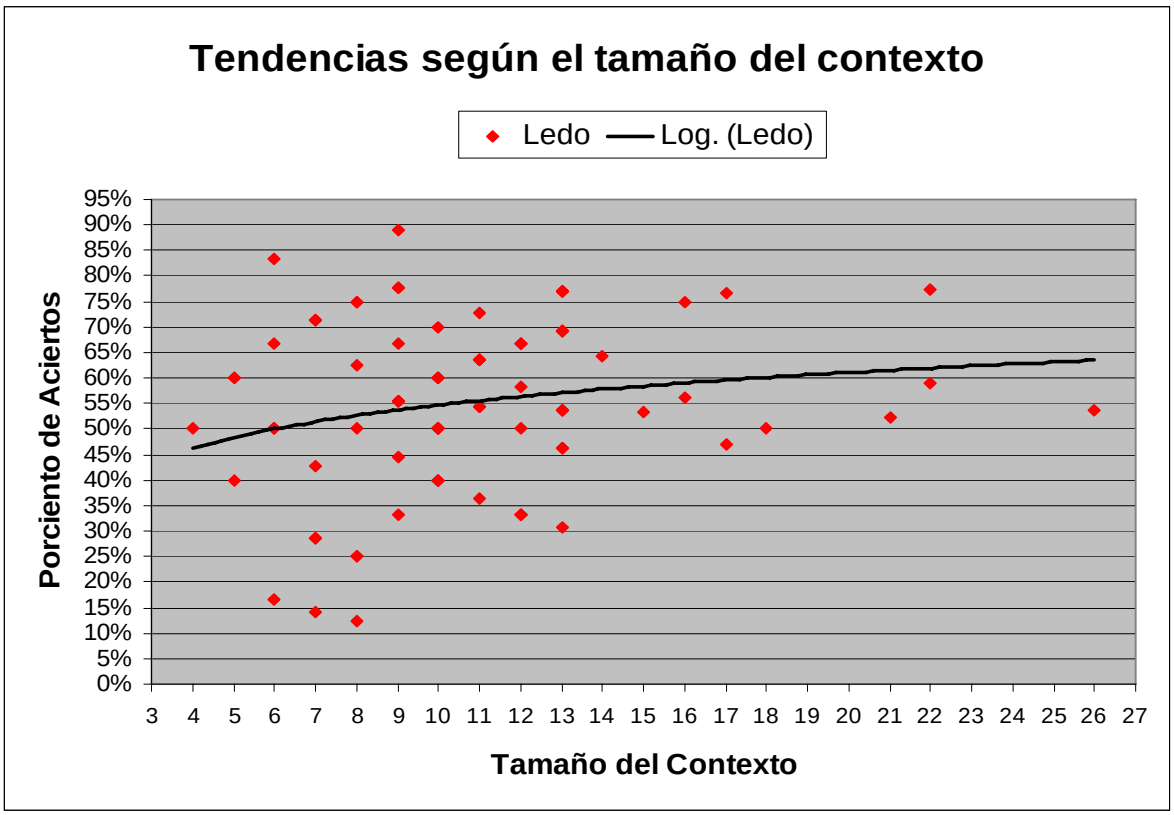

Ilustración 2. Porcentaje de aciertos para diferentes tamaños de contextos

Agradecimientos

Trabajo realizado con el apoyo parcial del gobierno de México (CONACyT, SNI) e Instituto Politécnico Nacional, Mexico (SIP, PIFI, COFAA).

Referencias

1. **Aguirre,** E. and G. **Rigau** (1996). Word Sense Disambiguation using Conceptual Density. Proc. 16th international conference on COLING. Copenhangen.
2. **Baeza-Yates**, R. and B. **Ribeiro-Neto** (1999). Modern Information Retrieval. Addison-Wesley.

3. **Bolshakov**, I. and A. **Gelbukh** (2004). Computational Linguistics: Models, Resources, Applications. IPN – UNAM – Fondo de Cultura Económica, Mexico, 186 p.
4. **Campos**, L. M. de (2001). Un modelo de recuperación de información basado en redes bayesianas. Universidad de Granada, España.
5. **Dolan**, W., L. **Vanderwende**, and S. **Richardson** (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.
6. **Ghazfan**, (1996). Toward meaningful Bayesian networks for information retrieval systems. In Proceedings of the IPMU´96 Conference, pages 841-846.
7. **Lesk**, M. (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In: Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June 1986, 24–26.
8. **Manning,** C. and H. **Schütze (**1999) Foundations of Statistical Natural Language Processing. MIT Press.
9. **Cowie,** J., L. **Guthrie**, and G. **Guthrie** (1992). Lexical disambiguation using simulated annealing. Proceedings of Coling-92, Nante, France, pp. 359-365.
10. **Global Reach** (2002). http://global-reach.biz
11. **McHale,** M. L. (1997). A comparison of WordNet and Roget's taxonomy for measuring semantic similarity.
12. **Lawrence**, S. (2000). El Acceso a la Información en la Web Limitado y Desigual. NEC Research Institute, http://www.neci.nec.com/
13. **McRoy**, S. (1992). Using multiple knowledge sources for word sense disambiguation. Computational Linguistics, Vol. 18(1), pp. 1-30.
14. **Mihalcea,** R. and D. **Moldovan** (1999). A Method for word sense disambiguation of unrestricted text. Proc 37th Annual Meeting of the ACL 152-158, Maryland, USA.
15. **Montoyo**, A. (2001). Método badaso en Marcas de Especificidad para WSD, Grupo de Procesamiento del Lenguaje y Sistemas de Información. Universidad de Alicante, España.
16. **Ravin**, Ya. and C. **Leacock** (2000). Polysemy: an overview. In Polysemy: Theoretical and Computational Approaches. Ravin Yael and Leacock Claudia (ed.). Oxford University Press. New York. 1-29
17. **Pimienta**, D. (2000). Representación de las lenguas y culturas latinas en la Internet, Fundación Redes y Desarrollo. Encuentro Sociedad y Tecnología, Santiago de Chile.
18. **Resnik**, Ph. (1995). Disambiguating noun groupings with respect to WordNet senses. Proc. Third Workshop on Very Large Corpora. 54-68. Cambridge, MA
19. **Resnik,** Ph. (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.
20. **Ribeiro**, B. (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
21. **Rigau,** G., J. **Atserias** and E. **Aguirre** (1997). Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. Proc 35th annual Meeting of the ACL, 48-55, Madrid, Spain.
22. **Saracevic**, T. (1995). A taxonomy of values for library and information services. Rutgers University, New Brunswick.
23. **Stetina** J., S. **Kurohashi** and M. **Nagao** (1998.) General word sense disambiguation method based on full sentencial context. In Usage of WordNet in Natural Language Processing. COLING-ACL Workshop, Montreal, Canada
24. **Sussna**, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. Proc. Second International CIKM, 67-74, Airlington.
25. **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.
26. **Voorhees**, E. 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.
27. **Wiks,** Y., D. **Fass**, C. **Guo**, J. **McDonal**, T. **Plate** and B. **Slator** (1993). Providing Machine Tractable dictionary tools. In: Semantics and the lexicon (J. Pustejowsky, Ed.), pp. 341-401
28. **Wilks**, Y. and M. **Stevenson** (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.
29. **Wilks,** Y. and M. **Stevenson**. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield, 1996.

30. **Wilks**, Y. and M. **Stevenson** (1998), Word sense disambiguation using optimized combination of knowledge sources. Proceedings of ACL 36/Coling 17, 1398-1402.
31. **WordNet**: an electronic lexical database. (1998), C. Fellbaum (ed.), MIT, 423 p.
32. **Yarowksy**, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceeding of Coling-92, Nante, France, pp. 454-460.

Yoel Ledo Mezquita *was born in La Habana, Cuba, in 1973. He obtained his Master degree in Electrical Engineering (Computer Science) in 1987 and his Ph.D. degree in Computer Science in 2006, from CIC-IPN, under supervision of Grigori Sidorov and Alexander Gelbukh. Since 2006, he works for the University of the Americas, Mexico. He is laureate of Lázaro Cardenas Prize of National Polytechnic Institute, México.*

Grigori Sidorov *was born in Moscow, Russia, in 1965. He obtained his Master degree in Structural and Applied Linguistics in 1988 from the Philological faculty of the "Lomonosov" Moscow State University, Russia, and his Ph.D. degree in Structural, Applied and Mathematical Linguistics in 1996 from the same faculty. Since 1998, he works for the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is a National Researcher of Mexico (SNI) since 1999; author of about 100 publications on computational linguistics; see www.cic.ipn.mx/~sidorov.*

Alexander Gelbukh *was born in Moscow, Russia, in 1962. He obtained his Master degree in Mathematics in 1990 from the department of Mechanics and Mathematics of the "Lomonosov" Moscow State University, Russia, and his Ph.D. degree in Computer Science in 1995 from the All-Russian Institute of the Scientific and Technical Information (VINITI), Russia. Since 1997, he is the head of the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is an academician of Mexican Academy of Sciences since 2000 and National Researcher of Mexico (SNI) since 1998; author of about 300 publications on computational linguistics; see www.Gelbukh.com.*