Automatización y evaluación de un método basado en conocimiento para la desambiguación en Mapas Conceptuales

Implementation and evaluation of a knowledge-based method for Concept Maps disambiguation

Manuel de la Iglesia Campos, Alfredo J. Simón Cuevas, Amhed Suárez Rodríguez, Alejandro Rosete Suárez

Instituto Superior Politécnico José Antonio Echeverría.

miglesia@ceis.cujae.edu.cu

Resumen

Los mapas conceptuales constituyen una herramienta eficiente para la representación del conocimiento en lenguaje natural. Un aspecto importante para su procesamiento de forma automatizada o semi-automatizada está dado por la determinación del sentido correcto, o desambiguación, de los conceptos de un mapa conceptual. El método de Desambiguación del Sentido de los Conceptos en mapas conceptuales, CSD, se basa en el análisis contextual, de glosa y dominio, utilizando WordNet como repositorio de sentidos. En este trabajo se presenta la automatización, mejora y evaluación del método CSD, y la comparación del método con el estado del arte, empleando un WordNet en español, y seleccionando un grupo de mapas conceptuales en dicho idioma como conjunto de evaluación. La versión final del método obtuvo en las pruebas realizadas 86.7% de precisión, 85.7% de exactitud y 98.8% de cobertura, resultados que pueden calificarse de alentadores comparados con el estado del arte.

Palabras clave: Desambiguación, mapas conceptuales, representación del conocimiento, sentidos.

Abstract

Concept maps are an efficient tool for knowledge representation in natural language. An important aspect for their automated or semi-automated processing is the determination of the right sense, or disambiguation, of the concepts in a concept map. The concept maps Concept Sense Disambiguation method, CSD, is based on the contextual, gloss and domain analysis, using WordNet as a sense repository. In this paper is presented the implementation, improvement and evaluation of the CSD method, and the comparison of the method with the state-of-the-art, using a Spanish WordNet and selecting a group of concept maps in that language as the experimentation set. The final version of the method obtained in the evaluation a precision of 86.7%, an accuracy of 85.7%, and coverage of 98.8%, results that can be qualified of encouraging in comparison with the state-of-the-art.

Key words: *Concept maps, disambiguation, knowledge representation, senses.*

Introducción

Los mapas conceptuales (MCs) constituyen una herramienta eficiente para la representación del conocimiento en lenguaje natural. Han sido definidos como "…una técnica que representa, simultáneamente, una estrategia de aprendizaje, un método para captar lo más significativo de un tema y un recurso esquemático para representar un conjunto de significados conceptuales incluidos en una estructura de proposiciones" (Novak y Gowin, 1984).

Los elementos estructurales, la facilidad de su construcción, así como la flexibilidad que brindan los MCs, los convierten en una herramienta ideal para la gestión del conocimiento y el aprendizaje en los humanos. No obstante, la propia flexibilidad que ofrecen los MCs constituye una limitante para la extracción del conocimiento de los mismos de forma automatizada, por lo que se hace necesario el empleo de técnicas y herramientas del Procesamiento del Lenguaje Natural (PLN) que permitan la extracción y utilización de este conocimiento (RAE, 2008). Entre estas técnicas y herramientas se encuentran las bases de conocimiento, los

analizadores morfológicos y la desambiguación. La desambiguación del sentido de las palabras (WSD, siglas del inglés Word Sense Disambiguation) consiste en determinar el sentido o significado de palabras ambiguas en un determinado contexto, y en el caso específico de los MCs puede ser aplicada para determinar el sentido correcto de los conceptos. Esto es posible ya que el MC puede ser considerado como un texto estructurado al estar conformado por proposiciones que asumen el papel de las oraciones del texto (Aguilar, 2004).

En tal sentido ha sido definido el procedimiento de Desambiguación del Sentido de Conceptos CSD, siglas del inglés *Concept Sense Disambiguation* (Simón et al., 2008), el cual permite la reducción de la ambigüedad de los conceptos presentes en un MC y se basa en el análisis contextual, de glosa y Dominios de Magnini (Magnini et al., 2002), utilizando WordNet (Miller et al., 1993) como repositorio de sentidos y aplicando transformaciones morfológicas a los conceptos que no se encuentren inicialmente en WordNet. Se han propuesto versiones anteriores del método en diferentes trabajos (Simón et al., 2006) (Simón et al., 2007). Debido a la complejidad en la ejecución del método en su versión actual, se hacía muy compleja la evaluación del mismo de forma manual. Se realizaron reducidas pruebas al método de desambiguación que permitieron comprobar en un nivel muy general su rendimiento y que podía ser aplicado a los MC. Sin embargo fue necesario crear las condiciones desde el punto de vista computacional que permitieran extender la evaluación del método, así como realizar dicha evaluación y permitir la integración del método en un contexto práctico.

En este trabajo se presenta la implementación y evaluación del método CSD, abarcando este último elemento la evaluación del método inicialmente de forma completa y posteriormente fragmentada en desambiguación por dominio, contexto y glosa; y la comparación con el estado del arte de la WSD en MCs.

Desarrollo

WSD

La ambigüedad puede ser definida como la característica de ofrecer más de una interpretación, y originar duda o confusión (Aristos, 1977); como un estado o condición equívoca que puede entenderse o interpretarse de varios modos, por ser poco clara o precisa; o como la condición de una palabra o frase que admite más de una interpretación (Larousse, 1998). En el lenguaje natural el pensamiento humano, de carácter abstracto, es representado mediante palabras o frases, lo que ocasiona que dichos elementos del lenguaje natural adquieran un significado determinado en dependencia del contexto en que se empleen. Esta ambigüedad presente en la mayoría de las palabras se denomina polisemia, que no es más que cuando una palabra puede adoptar varios sentidos o significados (RAE, 2008) (Larousse, 1998).

La desambiguación del sentido de las palabras (WSD, siglas del inglés Word Sense Disambiguation) es definida como el problema de seleccionar en un contexto determinado una definición o significado (sentido) para una palabra a partir de un conjunto de posibles sentidos de esa palabra (Pedersen y Mihalcea, 2005) (Pons, 2007) (Agirre y Edmonds, 2006). La WSD no constituye un fin en sí misma, sino que permite el desarrollo de otras tareas y aplicaciones de lingüística computacional y PLN tales como el análisis gramático, la corrección ortográfica, la interpretación semántica, la traducción computarizada, el análisis temático del contenido, la recuperación de información y la minería de datos (Pons, 2007).

Los métodos de WSD pueden clasificarse en tres grupos principales, según (Pedersen y Mihalcea, 2005), aunque no es una clasificación excluyente, pudiendo existir métodos que se puedan clasificar en más de un grupo. Estos grupos son:

Desambiguación basada en conocimiento: Consiste en el empleo de recursos léxicos externos tales como diccionarios [Collins (Collins, 2008) , Oxford (Oxford, 2002) y LDOCE (Longman, 2008) (Procter, 1978)], que incluyen una lista de significados, la

definición de cada uno de los términos y ejemplos típicos de su uso para la mayoría de ellos; tesauros [Tesauro Roget (Roget, 1952)], que agregan relaciones explícitas de sinonimia; y las redes semánticas [WordNet (Miller et al., 1993) y EuroWordNet (Vossen, 1996)] que añaden más relaciones semánticas.

Desambiguación supervisada: Consiste en un conjunto de métodos que inducen a un clasificador a partir de entradas etiquetadas manualmente y empleando el aprendizaje automático, basado en reunir un conjunto de ejemplos que ilustran las posibles clasificaciones o salidas de un evento, identificar patrones en los ejemplos asociados con cada clase particular del evento, generalizar estos patrones en reglas y aplicar las reglas para clasificar un nuevo evento (Pedersen y Mihalcea, 2005). Esta desambiguación emplea como recursos el texto etiquetado con los sentidos, un inventario de sentidos implícito (diccionario) y el análisis sintáctico de las palabras. Generalmente su alcance es de desambiguar una palabra por contexto, donde es necesario conocer la parte del habla de la palabra, así como formular explícitamente la palabra a desambiguar. Reduce la WSD a un proceso de clasificación de la palabra a desambiguar en uno de sus posibles sentidos.

Desambiguación no supervisada: Utiliza el aprendizaje no supervisado, consistente en identificar patrones en grandes cantidades de datos no etiquetados. Estos patrones se utilizan para dividir los datos en clústeres, donde cada miembro de un clúster tiene más en común con los otros miembros de su propio clúster que cualquiera de otro clúster diferente (Pedersen y Mihalcea, 2005). En este acercamiento a la desambiguación se emplea el aprendizaje no supervisado para encontrar similitudes entre contextos, y se desambigua sin necesidad de etiquetar los sentidos, sino sólo discriminando entre los clústeres y determinando cuáles ocurrencias tienen el mismo sentido sin necesidad de decir cuál es, por lo que también se le conoce a este tipo de método como discriminación.

Los MCs son representaciones simplificadas de la estructura cognitiva de una persona (o grupo de personas), en determinado dominio del conocimiento. Los conceptos, que constituyen representaciones abstractas de ideas, están altamente interconectados, formando proposiciones que expresan nuevas dimensiones de los conceptos originales, empleándose el lenguaje natural como principal herramienta para representar estos elementos, que por tanto están sujetos a ambigüedad (Valente et al., 2004). A su vez el MC puede ser considerado como un texto estructurado al estar conformado por proposiciones que asumen el papel de las oraciones del texto (Aguilar, 2004). Al ser posible analizar al MC como un texto estructurado es posible emplear la WSD para desambiguar los conceptos ambiguos del MC.

La información que se puede aprovechar en los MCs para la WSD es mucho más amplia que la existente en el texto plano. En primer lugar, la interconexión entre los nodos del MC se encuentra explícita, mientras que para el texto es necesario determinar las relaciones entre las palabras que conforman las oraciones. A diferencia de la definición de contexto en textos como son la oración, el párrafo y/o el documento completo, la definición del contexto de un concepto en un MC es mucho más precisa. El contexto de un concepto del MC está dado por la vecindad que se puede definir a partir de ese concepto y que está conformada por el conjunto de conceptos relacionados directamente con el concepto a diferentes niveles de profundidad.

El uso de los Dominios de Magnini es otro elemento que ayuda a reducir considerablemente la polisemia de los conceptos. Los MCs, por lo general, expresan un conocimiento de un dominio específico, guiado por el concepto principal, por lo que el uso de dominios para la desambiguación se encuentra muy relacionado con esto (Simón et al., 2006).

La glosa que describe a los synsets en WN puede utilizarse para reducir la ambigüedad de un concepto en los MCs, de forma tal que el sentido correcto del concepto debe ser con mayor probabilidad el que exhiba la mayor frecuencia de aparición de los conceptos del contexto en su glosa.

Método CSD

El método de Desambiguación del Sentido de los Conceptos CSD (Simón et al., 2008) tiene como entrada el MC y comprende cinco pasos:

Paso 1. Preparación del MC

Se extrae el conjunto de todos los conceptos y las proposiciones a las que pertenecen del MC, y se crean los siguientes conjuntos:

Conjunto de conceptos desambiguados

Conjunto de conceptos desconocidos

Conjunto de conceptos ambiguos

Paso 2. Selección de los dominios del MC

Son extraídos todos los Dominios de Magnini presentes en cada sentido asociado a cada concepto perteneciente al conjunto de todos los conceptos de una vecindad de radio r en la que el concepto raíz es su centro. De estos dominios se seleccionan como dominios principales del MC aquellos de mayor frecuencia y los dominios de los sentidos del concepto raíz que contengan uno de estos dominios más frecuentes.

Paso 3. Desambiguación por dominio

Los conceptos son desambiguados a partir de los dominios principales del MC. Se recorre cada uno de los conceptos no desambiguados, el concepto será desambiguado si solo uno de sus sentidos tiene un dominio que se encuentra en este conjunto.

Paso 4. Desambiguación por contexto

El sentido correcto de un concepto aún ambiguo será aquel a partir del cual se pueda crear una vecindad en WordNet en la que se encuentre la mayor cantidad de los sentidos correspondientes a los conceptos del contexto en el que se encuentra el concepto en el MC y estén a la menor distancia del centro de dicha vecindad. El contexto del concepto se define inicialmente a partir de aquellos conceptos a distancia igual o menor que 2 del concepto a desambiguar, esta distancia se incrementa iterativamente hasta hallar un sentido para desambiguar el concepto o hasta que el contexto abarque el total de conceptos del MC.

Paso 5. Desambiguación por glosa

En este paso se emplea la glosa que describe a los sentidos en la desambiguación de los conceptos ambiguos. Se determina la frecuencia de ocurrencia de los conceptos del contexto en que se encuentra el concepto a desambiguar en la glosa de los sentidos de dicho concepto, el que será desambiguado con aquel sentido, en caso de ser sólo uno, con mayor frecuencia de ocurrencia.

Si luego de este paso queda algún concepto sin desambiguar, entonces el método no propone ningún sentido para ese concepto.

Implementación del método CSD

Se decidió escoger como plataforma de desarrollo el IDE Eclipse 3.2, por su carácter de software libre y por las facilidades y flexibilidades que brindan este IDE y Java como lenguaje de programación. La herramienta implementada, nombrada Disambiguator, toma como entrada el MC en formato XML y devuelve el MC con los sentidos de los conceptos desambiguados en igual formato. Para facilitar su posible integración con otras herramientas como la plataforma GeCoSoft (Simón, 2006) y CmapTools (Cañas y Carvalho, 2004), se implementó como servicio web (de la Iglesia, 2008), empleando para ello las potencialidades para la creación de servicios web brindadas por MyEclipse Enterprise Workbench versión 5.0.1 GA, basadas en el framework XFire.

El modelo de dominio presentado en la Fig. 1 describe de forma general aquellos aspectos del sistema que son importantes en su contexto. La herramienta permite la desambiguación de los conceptos de un MC, por consiguiente los conceptos principales que

se modelan en el dominio están en función del MC y de los elementos necesarios para la desambiguación de los conceptos del mismo. Estos conceptos se describen a continuación:

- Conceptos representados en el MC que son procesados por la herramienta y pueden ser identificados con el nombre.
- Enlaces que unen a los conceptos.
- Base de Conocimientos que es empleada en el proceso de desambiguación.
- Términos que se encuentran presentes en la base de conocimientos.
- Sentidos asociados a cada término de la base de conocimiento.
- Dominios correspondientes a cada sentido.
- Tipos de Relación existentes entre los diferentes sentidos o synsets.
- MC en formato XML que es cargado por la herramienta.
- MC en formato XML enriquecido con los sentidos de cada concepto, obtenido como resultado de la desambiguación de los conceptos del MC.

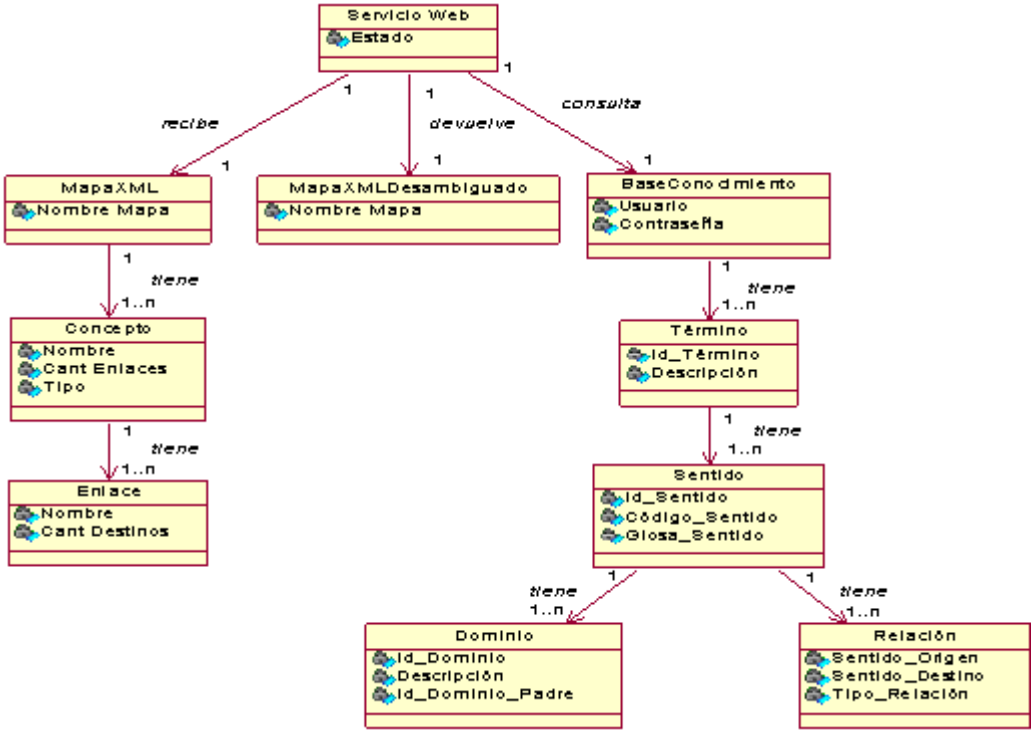

Fig. 1. Modelo de Dominio

En la Fig. 2 se presenta el diagrama de despliegue que modela un contexto en el que puede integrarse la herramienta, así como la interacción de dicha herramienta con WordNet.

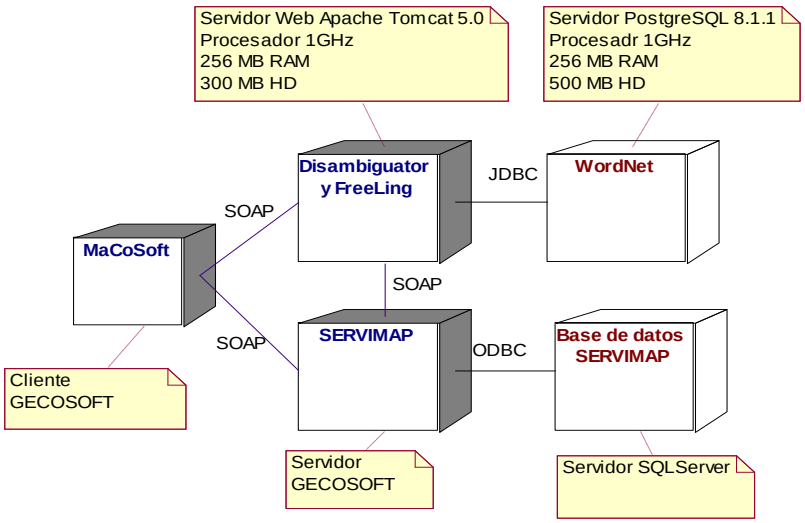

Fig. 2. Un contexto en el que puede integrarse la herramienta Disambiguator

WordNet

WordNet (Miller et al., 1993) es un sistema electrónico desarrollado en forma de base de datos léxica, creado en la Universidad de Princeton. Su estructura está basada en synsets (conjunto de uno o más sinónimos) representando cada uno un concepto léxico, que están etiquetados con un identificador y una glosa o descripción del significado, e interaccionan entre ellos por medio de diferentes tipos de relación, entre las que se encuentran hiperonimia/hiponimia, meronimia/holonimia, antonimia, causa-efecto, roles y sinonimia, cada una de ellas con sus acepciones y especificaciones. WordNet contiene una taxonomía de dominios de ámbito general, incorporados a partir de una propuesta de Magnini (Magnini et al., 2002). Cada synset tiene asociado uno o varios dominios. A partir de las relaciones léxicas presentes en WN puede ser identificada la semántica que expresa una relación entre dos términos, asociados a dos conceptos.

EuroWordNet (Vossen, 1996) representa una especie de extensión de WordNet que permite la interconexión de varios idiomas, integrándolos mediante Índices InterLingua (ILI) que permiten la obtención de conceptos semánticamente relacionados por relaciones definidas en los distintos idiomas.

El algoritmo CSD se clasifica como basado en conocimiento ya que usa WordNet como repositorio de sentidos; por lo tanto su aplicación en un idioma u otro solo requiere de usar la versión de WordNet correspondiente al idioma en cuestión. Se empleó como base de conocimientos un WN en idioma español incluido en una versión de EWN desarrollada en el Grupo de Procesamiento de Lenguaje Natural (TALP) del Departamento de Lenguajes y Sistemas Informáticos (LSI) de la Universidad Politécnica de Cataluña (UPC), cuya última versión data de 2006, siendo realizado a partir del WordNet 1.6 en inglés. Se realizó una caracterización de ambos WordNet, cuyos resultados se presentan en la Tabla 1.

Tabla 1. Caracterización de WordNet 1.6 en inglés y en español

Versiones de WordNet	CT[1]	CS[2]	CTA[3]	PSTA[4]	CTR[5]
WordNet v1.6 en Español	59 793	104 794	15 074	3,26	69
WordNet v1.6 en Inglés	129 509	99 642	23 255	2,90	19

[1] Cantidad de términos.
[2] Cantidad de synsets.
[3] Cantidad de términos ambiguos
[4] Promedio de synsets por término ambiguo.
[5] Cantidad de tipos de relaciones.

Mediante este estudio sobre la información almacenada en las versiones 1.6 de WordNet en inglés y español se llegó a las siguientes conclusiones:

El WordNet en español incluye un 53,8% menos términos que la versión en inglés.

Existe mayor proporción de términos ambiguos en WordNet en español (25,2%) que en inglés (18%).

Los conceptos ambiguos en WordNet en español tienen mayor ambigüedad (3,26 synset por concepto ambiguo como promedio) que en el inglés (2,90).

Adicionalmente se realizó un enriquecimiento del WordNet en español a partir de la versión en inglés, incorporándole específicamente las relaciones entre synsets presentes en el WordNet en inglés que no se encontraran en el WordNet en español, con lo que se logró la incorporación a este WordNet de 97887 nuevas relaciones, clasificadas como se describe en la tabla 2.

Tabla 2. Enriquecimiento del WordNet 1.6 en español a partir de WordNet 1.6 en inglés

Tipo de relación	Cantidad de relaciones agregadas
has_mero_member (meronimia)	11848
has_mero_madeof (meronimia)	710
has_hyperonym (hiperonimia)	78446
has_mero_part (meronimia)	6883
Total	**97887**

Experimentación

La vía utilizada para medir el rendimiento de CSD ha sido la aplicación experimental del método CSD sobre un conjunto de MCs en idioma español, utilizando un WordNet en el mismo idioma. Un inconveniente para este proceso es la no disponibilidad de un repositorio de MCs reconocido para este tipo de experimentos. Por este motivo fue necesario conformar un repositorio de MCs constituido por 30 MCs que agrupan un total de 545 conceptos, de ellos 409 presentes en WordNet y 251 conceptos ambiguos (con más de un synset en WordNet), teniendo como promedio 3.3 synsets por concepto y 5.03 synsets por concepto ambiguo. Se tomaron en consideración los siguientes criterios para la selección de los MC: debían ser representaciones correctas del conocimiento; el conocimiento representado en ellos debía corresponder a dominios poco profundos del conocimiento para evitar que los conceptos no se encontraran en WordNet; y cada MC debía tener al menos un concepto ambiguo para permitir la evaluación del método CSD. Las características detalladas del conjunto de estos mapas conceptuales son descritas en la tabla 3.

Tabla 3. Características generales del conjunto de experimentación

Nombre del MC	C[6]	CWN[7]	SCW[8]	CA[9]	SCA[10]	D[11]	DCA[12]
Agua II	22	21	3.62	13	5.23	24	21
Aire	9	8	5.63	5	8.40	14	12
Animales	21	21	2.04	10	3.20	12	10
Animales II	10	10	2.30	7	2.86	10	10
Atmósfera	17	10	2.70	3	8.33	8	6
Átomos	11	9	3.67	6	5.00	17	17

[6] Cantidad de conceptos del MC
[7] Cantidad de conceptos presentes en WordNet
[8] Cantidad de sentidos por concepto presente en WordNet**;**
[9] Cantidad de conceptos ambiguos
[10] Cantidad de sentidos por concepto ambiguo
[11] Cantidad de dominios por concepto
[12] Cantidad de dominios por concepto ambiguo.

Átomos-moléculas	23	17	3.29	11	5.00	23	23
Biología celular	28	11	1.91	2	6.00	9	8
Células	23	15	2.27	5	4.60	13	12
Cerebro Humano	14	11	4.73	8	6.13	15	15
Ciclo del carbono	28	14	2.93	10	3.70	19	19
Clima	14	8	2.88	7	3.14	11	11
Ecosistema	9	7	2.28	2	5.50	13	10
Entorno	33	28	4.00	24	4.50	29	29
Entrevista	38	24	3.00	19	3.58	30	29
Fotosíntesis	17	14	3.00	5	6.60	14	12
Fotosíntesis II	16	12	2.33	3	6.33	13	7
Geología	12	12	2.08	4	4.25	11	10
Humanismo	18	11	3.64	9	4.22	15	15
Magnitudes	19	14	4.14	13	4.38	18	18
Matemática	21	12	8.00	10	9.40	25	25
Minerales	7	4	2.50	2	4.00	6	5
Molécula	13	13	4.92	10	6.10	21	20
Nitrógeno	13	9	2.67	5	4.00	16	14
Nitrógeno II	22	18	4.44	11	6.36	21	19
Océano	13	8	2.13	3	4.00	13	10
Plantas	16	16	2.50	12	3.00	20	20
Ser Vivo	20	19	2.26	10	3.40	11	9
Universo	26	23	5.00	17	6.41	32	31
Vasos sanguíneos	12	10	2.10	5	3.20	14	13
Promedio	**18.17**	**13.63**	**3.3**	**8.37**	**5.03**	**16.57**	**15.33**
Desv. Estándar	**7.38**	**5.64**	**1.36**	**5.28**	**1.70**	**6.63**	**7.05**
Total	**545**	**409**		**251**			

El conjunto de MCs posee un total de 545 conceptos, el 75% de ellos presentes en WordNet, y de estos últimos 251 conceptos ambiguos (el 61% de los conceptos que están presentes en WordNet). El conjunto de evaluación se conformó con los 251 conceptos ambiguos representados en los MCs, los que estaban presentes en un promedio de cinco synsets, evidenciando el elevado nivel de granularidad de los sentidos en el WordNet utilizado.

El experimento se dividió en tres etapas: Aplicación del método CSD; Aplicación de la desambiguación por dominio, contexto y glosa de CSD de forma independiente; y Comparación de CSD con el algoritmo reportado por Cañas y colaboradores (Cañas et al., 2003). Para medir los resultados de las pruebas fueron utilizadas las métricas de precisión (precisión), exactitud (recall o accuracy) y cobertura (coverage), métricas muy utilizadas en la evaluación de algoritmos de WSD. El cómputo de los resultados se realizó partiendo de conocer el sentido esperado de cada concepto a desambiguar y comparándolo con la respuesta obtenida por el algoritmo.

Aplicación del método CSD

La tabla 4 muestra los resultados de la aplicación del método CSD de forma completa. Se han omitido los resultados específicos obtenidos en cada MC. De igual forma se presentan los resultados en las tablas 5, 6, 7 y 8.

Tabla 4. Resultados de la desambiguación aplicando CSD

	CDC	CDI	CND	CD	P	E	C
Promedio	2.80	0.23	5.33	3.03	0.914	0.370	0.399
Desviación Estándar	2.295	0.430	4.097	2.442	0.213	0.215	0.217
Total	**84**	**7**	**160**	**91**	**0.933**	**0.336**	**0.360**

Tabla 5. Resultados de la desambiguación por dominio

	CDC[13]	CDI[14]	CND[15]	CD[16]	P[17]	E[18]	C[19]
Promedio	7.17	1.1	0.1	2.77	0.856	0.842	0.979
Desviación Estándar	4.68	1.37	0.31	2.44	0.145	0.166	0.071
Total	**215**	**33**	**3**	**248**	**0.867**	**0.857**	**0.988**

Se desambiguaron 91 conceptos (36%) usando la desambiguación por dominio, 157 conceptos (62,5%) usando la desambiguación por contexto y no fue necesario el uso de la desambiguación por glosa. El algoritmo obtuvo un 86,7% de precisión, un 85,7% de exactitud y un 98.8% de cobertura, como resultados generales.

Los resultados significativamente inferiores al promedio en algunos MCs se debieron a la combinación de los siguientes factores:

Pocos conceptos incluidos en algún synset en WordNet: existen nueve MCs con menos del 65% de sus conceptos presentes en algún synset en WordNet.

Elevada ambigüedad de los conceptos ambiguos: los conceptos ambiguos representados en los MCs están presentes en más de cinco synsets como promedio.

Aplicación de los pasos del método CSD

Los pasos esenciales del método: la desambiguación por dominio, contexto y glosa, se evaluaron de forma independiente con el objetivo de comprobar su incidencia en la desambiguación. Los resultados de la desambiguación por dominio, contexto y glosa se presentan en las tablas 5, 6 y 7 respectivamente.

Tabla 6. Resultados de la desambiguación por contexto

	CDC	CDI	CND	CD	P	E	C
Promedio	6.93	1.33	0.10	8.27	0.845	0.827	0.979
Desviación Estándar	4.608	1.470	0	5.349	0.152	0.160	0.071
Total	**208**	**40**	**3**	**248**	**0.839**	**0.829**	**0.988**

Tabla 7. Resultados de la desambiguación por glosa

	CDC	CDI	CND	CD	P	E	C
Promedio	1.70	0.50	6.17	2.2	0.807	0.194	0.264
Desviación Estándar	1.745	1.042	4.511	2.384	0.286	0.179	0.231
Total	**51**	**15**	**185**	**66**	**0.773**	**0.203**	**0.263**

[13] Conceptos desambiguados correctamente
[14] Conceptos desambiguados incorrectamente
[15] Conceptos no desambiguados
[16] Conceptos desambiguados
[17] Precisión
[18] Exactitud
[19] Cobertura

Como principales resultados se destacan la alta precisión (93%) obtenida con la desambiguación por dominio, la alta cobertura presentada en la desambiguación por contexto (99%), que incide en la cobertura presentada por el método CSD de forma general, y los bajos resultados de exactitud y cobertura en la desambiguación por dominio (33,3% y 36% respectivamente) y por glosa (exactitud: 20,3% y cobertura: 26,2%). Las causas de estos bajos resultados en específico son:

Un mismo dominio puede agrupar diferentes synsets de una misma palabra; estos casos no pueden ser desambiguados.

En WordNet 1.6 en español la existencia de glosa en los synsets es muy baja, el 28% de los conceptos ambiguos procesados no tienen ningún synset con glosa y las que existen incluyen pocas palabras, lo que afecta también la precisión.

La precisión obtenida en cada uno de los pasos de forma independiente justifica el orden de los pasos en el algoritmo, un orden de mayor a menor precisión (desambiguación por dominio, luego por contexto y finalmente por glosa) lo que permite alcanzar un resultado de precisión óptimo para el método CSD completo.

Comparación del método CSD con algoritmo de cañas y colaboradores

Se decidió comparar el método CSD con el algoritmo de Cañas y colaboradores (Cañas et al., 2003) debido a que es el único método conocido que, al igual que CSD, realiza la desambiguación de conceptos en MCs. Para esto fue necesario implementar dicho algoritmo tal y como está definido por sus autores (Cañas et al., 2003). La tabla 8 muestra de manera general los resultados obtenidos por cada algoritmo empleando el mismo conjunto de experimentación.

Tabla 8. Comparación entre CSD y el algoritmo de Cañas y colaboradores

	CSD						Cañas y colaboradores					
	CDC	CDI	CND	P	E	C	CDC	CDI	CND	P	E	C
Promedio	7.17	1.1	0.1	0.86	0.84	0.98	4.93	3.47	0	0.51	0.51	1.00
Desv. Est.	4.68	1.37	0.31	0.15	0.17	0.07	4.28	1.98	0	0.26	0.26	0
Total	**215**	**33**	**3**	**0.87**	**0.86**	**0.99**	**148**	**104**	**0**	**0.59**	**0.59**	**1.00**

Empleando el método CSD se obtuvo un 87% de precisión y 86% de exactitud, y un 59% de precisión y exactitud utilizando el algoritmo de Cañas. La cobertura fue de 99% y 100%, respectivamente.

Resultados y discusiones

Un problema presentado en el proceso de evaluación es la granularidad muy fina de los sentidos en WordNet, lo que dificulta la determinación del sentido correcto de un término en un contexto dado.

El método CSD mostró como resultados generales un 0.87 de precisión, 0.86 de exactitud y 0.99 de cobertura, los que pueden ser evaluados de positivos y alentadores, incluso dentro del contexto de los algoritmos de WSD en general, y de los MC en particular.

La utilización en la desambiguación de tipos de conocimiento variados tales como los dominios, las relaciones de hiperonimia/hiponimia, holonimia/meronimia y glosa, y la glosa que describe a los sentidos en WordNet; y de fuentes de conocimientos diversas, como WordNet, tesauros y ontologías, son elementos que se pueden aprovechar para obtener algoritmos de WSD eficientes.

Entre los elementos que influyeron en los resultados obtenidos de manera general se encuentran:

La no inclusión en CSD de mecanismos para el tratamiento de conceptos que no estén presentes en algún synset en WordNet, lo que constituye una limitante del método.

El insuficiente desarrollo de la versión en español de WordNet, que dificulta la obtención de mejores resultados en los procesos de desambiguación.

La no existencia de un repositorio de MCs disponible para la realización de este tipo de experimentos, lo que conllevó a la selección de MCs que representan conocimientos muy específico; sin embargo, WordNet almacena conocimiento de dominio general

Trabajos futuros

Los términos y sus sentidos, entre otros elementos, varían en dependencia del idioma, por lo que evaluar el método CSD en un idioma diferente al español puede arrojar resultados diferentes a los obtenidos. En esto tiene un papel fundamental la composición del WordNet a utilizar. Resulta necesario por tanto extender la evaluación del método al idioma inglés, lo que posibilitará un mejor análisis de la influencia del idioma y de los elementos que integran WordNet en la desambiguación, específicamente en el método CSD.

Deben incorporarse a CSD mecanismos para usar sinónimos de conceptos (que no aparecen en WordNet) en fuentes externas; por ejemplo, un corpus de MCs. Otro elemento a realizar en futuros trabajos es la creación de un repositorio de MCs de dominios generales de conocimiento.

Conclusiones

Se ha presentado la implementación y evaluación del método CSD de desambiguación del sentido de los conceptos en un MC. Los resultados obtenidos por CSD en la experimentación pueden ser evaluados de positivos, alcanzando valores de 87% de precisión, 86% de exactitud y 99% de cobertura. La precisión obtenida de forma independiente por los pasos de desambiguación por dominio, contexto y glosa justifican el orden de estos pasos en el método, pues se ejecutan de mayor a menor precisión, permitiendo alcanzar con esto un resultado de precisión óptimo para el método CSD completo.

Referencias Bibliográficas

[1] (Aguirre y Edmonds, 2006) E. Agirre y P. Edmonds, "Word Sense Disambiguation: Algorithms and Applications". Springer, 2006. URL: http://wsdbook.org/index.html

[2] (Aguilar, 2004) M.F. Aguilar, "El Mapa Conceptual: un texto a interpretar". Proceedings of the First International Conference on Concept Mapping 2004 (CMC'04), Pamplona, España, 2004.

[3] (Aristos, 1977) Diccionario Ilustrado de la Lengua Española Aristos. Editorial Científico-Técnica, Ciudad de la Habana, 1977.

[4] (Cañas et al., 2003) A.J. Cañas, A. Valerio, J. Lalinde Pulido, M. Carvalho y M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction". En Proceedings of 10th International Symposium on String Processing and Information Retrieval. Springer-Verland. Manaus, Brasil, 2003.

[5] (Cañas y Carvalho, 2004) A.J. Cañas, y M. Carvalho, "Concept Maps and AI: an Unlikely Marriage?" Institute for Human & Machine Cognition, Pensacola, FL 32502, 2004.

[6] (Collins, 2008) Collins Dictionaries. Harper Collins Publishers. URL: http://www.collinsdictionaries.com. Consultado en Mayo, 08 de 2008.

[7] (de la Iglesia, 2008) M. de la Iglesia, "Automatización del proceso de desambiguación de conceptos en mapas conceptuales: Disambiguator". Trabajo de diploma para optar por el titulo de Ingeniería Informática, Instituto Superior Politécnico "José Antonio Echeverría", Ciudad de La Habana, 2008.

[8] (Larousse, 1998) Gran Diccionario de la Lengua Española Larousse. Larousse Editorial S.A., 1998. ISBN: 84-816-266-X.

[9] (Longman, 2008) Longman Dictionary of Contemporary English (LDOCE). URL: http://www.longman.com/ldoce. Consultado en Mayo, 11 de 2008.

[10] (Magnini et al., 2002) B. Magnini, C. Strapparava, G. Pezzulo, y A. Gliozzo, "The Role of Domain Information in Word Sense Disambiguation". Natural Language Engineering. Cambridge University Press, 2002, pp. 359–373.

[11] (Miller et al., 1993) G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross y K. Miller, "Introduction to WordNet: an On-line Lexical Database". 1993.

[12] (Novak y Gowin, 1984) J.D. Novak y D.B. Gowin, "Learning how to learn". New York, NY: Cambridge, University Press, 1984.

[13] (Oxford, 2002) Oxford English Dictionary (OED). Segunda Edición, 2002. URL: http://www.oed.com

[14] (Pedersen y Mihalcea, 2005) T. Pedersen y R. Mihalcea, "Advances in Word Sense Disambiguation", Tutorial de la Association for Computational Linguistics (ACL), 2005.

[15] (Pons, 2007) A. Pons Porrata, "Una Panorámica a la Desambiguación del Sentido de las Palabras". Centro de Estudios de Reconocimiento de Patrones y Minería de Datos, RECPAT, 2007.

[16] (Procter, 1978) P. Procter, Longman Dictionary of Contemporary English. Longman, London, 1978.

[17] (RAE, 2008) Diccionario de la lengua española. Real Academia Española. Vigésimosegunda edición, URL: http://www.rae.es. Fecha de consulta: Mayo 7, 2008.

[18] (Roget, 1952) P.M. Roget "ROGET's THESAURUS of English Words and Phrases". 1952. URL: http://thesaurus.reference.com/

[19] (Simón, 2006) A.J. Simón Cuevas, "GECOSOFT: Plataforma para la Gestión del Conocimiento con Mapas Conceptuales" Tesis presentada en opción al Título de Máster en Informática Aplicada. Tutores: V. Estrada Sentí y A. Rosete Suárez. Instituto Superior Politécnico "José Antonio Echeverría", Ciudad de La Habana, 2006.

[20] (Simón et al., 2006) A.J. Simón, C. Ceccaroni, S. Willmott, A. Rosete, V. Estrada, y V. Lara "Modelo Unificado de Representación del Conocimiento en Mapas Conceptuales y Ontologías". Proceedings of the Second International Conference on Concept Mapping 2006 (CMC'06), Septiembre de 2006. URL: http://cmc.ihmc.us/papers/cmc2006-p153.pdf

[21] (Simón et al., 2007) A. Simón, L. Ceccaroni y A. Rosete, "Generation of OWL Ontologies from Concept Maps in Shallow Domains", CAEPIA 2007, LNAI 4788, Springer-Verlag, 259-267, 2007b.

[22] (Simón et al., 2008) A.J. Simón, L. Ceccaroni, A. Rosete, A. Suárez y M. de la Iglesia, "A Concept Sense Disambiguation Algorithm for Concept Maps". Proceedings of the Third International Conference on Concept Mapping 2008 (CMC'08), 2008.

[23] (Valente et al., 2004) J. Valente, F.E. Lopes y E. Luiz, "Linking Phrases in Concept Maps: A Study on the Nature of Inclusivity". Primer Congreso Mundial de Mapas Conceptuales. Pamplona, España, 2004.

[24] (Vossen, 1996) P. Vossen, "EuroWordNet: a multilingual database for information retrieval". Proceeding of the DELOS Workshop on Cross-Language Information Retrieval. Zurich, Switzerland, Marzo de 1997. URL: http://www.ercim.org/publication/ws-proceedings/DELOS3/Vossen.pdf