Procesamiento del Lenguaje Natural, num. 29 (2002), pp.231-238 recibido 25-05-2002; aceptado 03-07-2002

Representación conceptual basada en técnicas lingüísticas*

P. Moreda y R. Muñoz
Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Campus San Vicente del Raspeig
{moreda,rafael}@dlsi.ua.es

Resumen: Cada vez más, el acceso a la Sociedad de la Información requiere recursos y herramientas con mayores capacidades lingüísticas, y en especial con capacidades semánticas y conceptuales. Siguiendo esta línea de exposición, en este artículo se presenta una propuesta de representación conceptual basada en técnicas lingüísticas. Desde este punto de vista, en el artículo se plantean tanto problemas como propuestas de solución de representación de un concepto, de representación conceptual de un documento o de la extracción automática de patrones conceptuales. Además, en el trabajo se proponen las aplicaciones más inmediatas que se pueden llevar a cabo en una propuesta de estas características.
Palabras clave: Representación conceptual, semántica, WordNet

Abstract: The access to the Society of the Information requires More and more resources and tools with more linguistic capabilities, and especially with semantic and conceptual capabilities. Following this line of explanation, this paper presents a proposal of conceptual representation based on linguistic technicals. From this point of view, this paper proposes problems and solutions in order to represent a concept, represent a document conceptually or extract conceptual patterns automatically. Also this paper proposes the most immediate applications in order to obtain a proposal of this characteristics.
Keywords: Conceptual representation, semantic, WordNet

1 *Introducción*

Partiendo de que el concepto es la idea que concibe el entendimiento, este artículo se centra en la hipótesis de que dado un texto se puede obtener una representación conceptual automática del mismo mediante el uso de herramientas y recursos lingüísticos.

Para ello, los principales problemas que se abordan a lo largo de este trabajo son:

- ¿Cómo representar conceptualmente un texto?
- ¿Cómo representar conceptualmente un concepto?
- ¿Qué recursos léxicos y/o herramientas son necesarias para llevar a cabo un enfoque conceptual?
- ¿Qué aplicaciones se beneficiarían de una enfoque de estas características?

Un aspecto relevante y determinante en el proceso de modelización conceptual es la ambigüedad léxica en dominios no restringidos semánticamente, sin embargo algunos trabajos y proyectos están direccionando sus actividades al enriquecimiento de, por ejemplo, EuroWordNet con dominios semánticos concretos. Desde este punto de vista el proyecto Euroterm[1] tiene el objetivo de exten-

* Este trabajo esta subvencionado por la Comisión Interministerial de Ciencia y Tecnología (CICyT) mediante el proyecto *TUSIR* de referencia TIC2000-0664-C02/02 y por la Comisión Europea mediante el proyecto *EuroTerm* de referencia EDC-2214

[1] EuroTerm es un proyecto financiado por la Comisión Europea (EDC-2214) con una duración total de dieciocho meses (del 01/01/01 al 30/06/02) e incluido en las acciones preparatorias del programa *e-content*. El consorcio está formado por investigadores de las universidades de Patras (Grecia), de Tilburg (Holanda) y de Alicante (España). Los participantes españoles son miembros del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante. La información relativa a este proyecto puede encontrarse en http://dblab.upatras.gr.

ISSN 1135-5948

der WordNet con terminología medioambiental en los idiomas griego, holandés y español. EuroWordNet es una base de datos multilingual formada por WordNets genéricos en ocho idiomas europeos (Vossen, 1998). Los WordNets individuales incorporados en la base de datos central forman redes semánticas autónomas conectadas entre sí a través del índice inter-lenguas (*Inter-Lingual-Index* o ILI).

La utilización de este tipo de recursos es básica a la hora de aplicar técnicas lingüísticas relacionadas con la información semántica, como se verá en la sección 2 que aborda el problema de la influencia de las técnicas lingüísticas en la modelización conceptual. Posteriormente, en la sección 3, se muestra nuestra propuesta de representación conceptual automática basada en el uso de técnicas lingüísticas, comparando dicha propuesta con otras existentes. Además, se mostrará un pequeño caso de estudio en el que se aplicará nuestra propuesta de representación conceptual a la modelización de sistemas de información. Finalmente, se mostrarán las conclusiones obtenidas con este trabajo, así como una serie de líneas a desarrollar en un futuro.

2 *Influencia de las técnicas lingüísticas en la Representación Conceptual*

La modelización conceptual de un sistema de información consiste en el estudio de un documento o texto con el objetivo de identificar partes relevantes del texto y las posibles relaciones entre estas partes relevantes. Las técnicas lingüísticas pueden ser aplicadas con el objetivo de procesar el texto, añadir la información necesaria e identificar dichas partes relevantes y sus relaciones, lo cual puede ayudar a la realización automática del modelado conceptual de un texto o de un concepto. A continuación se muestran las principales técnicas o recursos que pueden usarse para dicha modelización:

- *Etiquetado léxico.*
- *Recursos semánticos.*
- *Desambiguación del sentido de las palabras.*
- *Resolución de la ambigüedad referencial.*
- *Análisis sintáctico.*

Las siguientes secciones explican de forma detallada cada una de estas técnicas.

2.1 Etiquetado léxico

El etiquetado léxico del texto objeto de estudio proporciona información que resulta vital para el desarrollo de esta modelización automática. Las entidades u objetos que aparecen al modelar conceptualmente un sistema de información serán nombres comunes, aunque no todos los nombres comunes serán entidades. Algunos nombres comunes podrán ser propiedades de una entidad y otros no tienen que considerarse al realizar la modelización conceptual. Las propiedades son partes de las entidades que las caracterizan y pueden hacerlas diferentes de otras entidades del mismo tipo. Mediante este etiqueta léxica se obtiene la información relativa a las categorías gramaticales de cada palabra, así como su raíz o lema. Para este trabajo se ha utilizado el etiquetador morfológico MACO (Acebo et al., 1994) con el desambiguador léxico desarrollado por Pla (Pla, 2000).

2.2 Recursos semánticos

El papel que juega la semántica en la mayoría de tareas relacionadas con el procesamiento de textos escritos en lenguaje natural es primordial. A su vez, la incorporación de este tipo de información añade un nuevo problema como es la desambiguación del sentido de las palabras. Dicha desambiguación proporciona el sentido correcto que tienen las palabras polisémicas en el contexto en el que se producen. La información semántica que se requiere para la modelización conceptual de textos escritos en lenguaje natural es, por un lado, las diferentes redes de sinonimia, hiperonimia, hiponimia, etc. de una determinada palabra (la figura 1 muestra la red asociada a la palabra *planta*) y, por otro lado, la utilización de una ontología que nos permita tener una estructura del conocimiento en diversas categorías para poder situar todas los objetos en las categorías apropiadas de dicha estructura (la palabra *planta* pertenece al conjunto de categorías 1stOrderEntity 20 Form Living Natural Object Origin Part Plant de la ontología de EuroWordNet que se muestra en la figura 2).

La aplicación de estos recursos semánticos a los textos objeto de estudio nos permite la extracción de patrones de conocimiento relacionando las categorías que acompañan a un

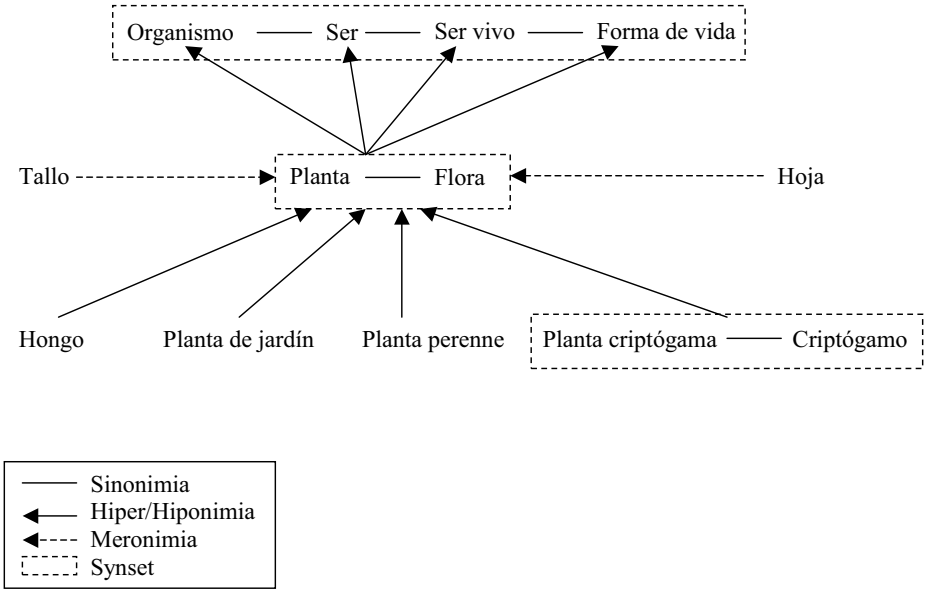

Figura 1: *Red semántica asociada a la palabra* Planta

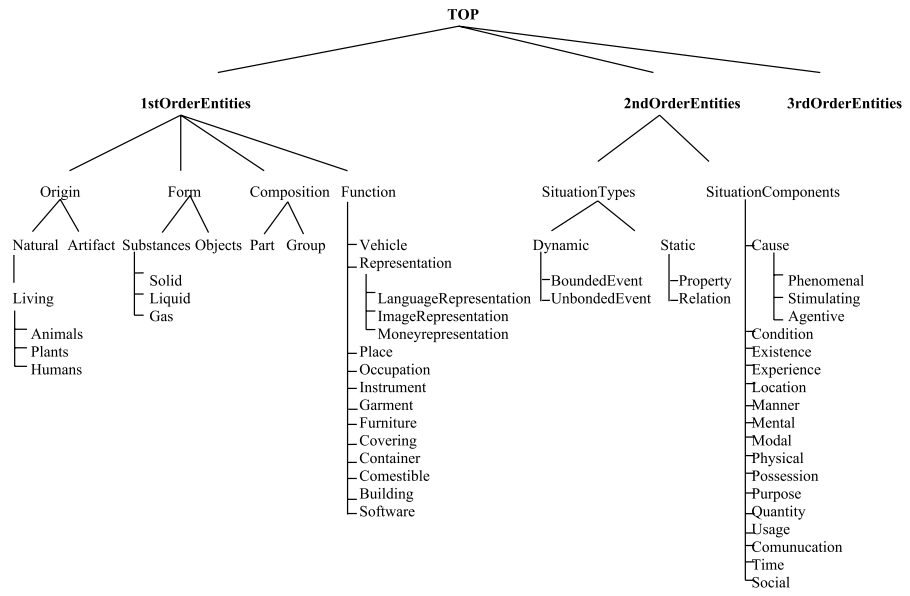

Figura 2: *Ontología de EuroWordNet*

conjunto de verbos característicos.

2.3 Resolución de la ambigüedad referencial

La escritura de un texto, en general, conlleva la utilización de un conjunto de técnicas que permiten mantener la coherencia del texto. El mecanismo normalmente usado para mantener dicha coherencia es la ambigüedad referencial (anáfora). La resolución de esta ambigüedad referencial que se produce en el texto es necesaria para poder detectar la aparición de entidades y de sus propiedades a lo largo del texto con distinto nombre o la aparición de forma abreviada de la misma con la utilización de pronombres. Además, esta resolución de la ambigüedad referencial proporciona información implícita que permite establecer la posible relación entre dos nombres de forma que uno describa a otro. Los algoritmos utilizados para la resolución de la ambigüedad referencial producido por pronombres y descripciones definidas son los desarrollados en (Palomar et al., 2001; Muñoz and Palomar, 2001), respectivamente. Estos algoritmos añaden al texto la información de las

cadenas de correferencias existentes de forma que se pueden establecer las relaciones existentes entre las diversas apariciones de las misma entidad a lo largo del texto.

2.4 Análisis sintáctico

La información acerca de la agrupación de las palabras en los distintos sintagmas (nominal, preposicional y verbal) dentro de una frase y el papel sintáctico que cada uno ellos juega en una determinada frase es primordial a la hora de poder establecer patrones de conocimiento. El análisis sintáctico proporciona toda esta información. En este trabajo se ha utilizado el analizador SUPP desarrollado en (Palomar et al., 1999). Este analizador proporciona información de los roles sintácticos (sujeto, objeto directo, etc.) que desempeñan en la frase a través de un conjunto de heurísticas.

3 Representación conceptual

3.1 Propuesta de representación conceptual

El objetivo que se persigue en este artículo es la investigación en representación conceptual de un texto o documento basada en técnicas lingüísticas.

La representación conceptual de un texto pasa por la representación conceptual de cada uno de los conceptos de dicho texto y de las relaciones entre ellos. Por ello, el primer paso necesario para obtener una representación conceptual de un texto será la identificación de los conceptos (subsección 3.1.1) que dicho texto incluya. Consideramos un concepto como la palabra o conjunto de palabras que constituyen el significado correcto del mismo junto con su información semántica asociada obtenida de una base de datos léxica.

Una vez identificados los conceptos contenidos en un texto, se deberá extraer las relaciones existentes entre ellos, así como los roles conceptuales de cada uno de ellos (objetos, relaciones o propiedades). Para ello, se hará uso de patrones conceptuales que permitan obtener de forma automática el modelo conceptual correspondiente a la información incluida en el texto. Estos patrones determinarán, por ejemplo, si un concepto describe a otro ((nombre es una propiedad que describe a las plagas), o si existe una determinada relación entre dos conceptos (las plagas atacan a los países), etc.

El proceso planteado presenta cierta similitud con la metodología utilizada en la herramienta LIDA (Overmyer, Lavoie, and Rambow, 2001), la cuál proporciona un apoyo lingüístico al proceso de desarrollo de modelos conceptuales. Sin embargo, mejora dicha herramienta en dos aspectos. Por un lado, el establecimiento de las relaciones entre conceptos es en LIDA un proceso totalmente manual llevado a cabo por analistas, mientras que el proceso presentado en nuestro trabajo determina de forma automática estas relaciones mediante el uso de patrones conceptuales tal y como se verá en la subsección 3.1.2. Por otro lado, en el proceso de obtención de la lista de conceptos de un texto en LIDA no se tienen en consideración las relaciones de sinonimia, no se hace desambiguación del sentido de las palabras, ni se resuelve la ambigüedad referencial, todas ellas técnicas lingüísticas que como se verá más adelante son utilizadas en nuestro proceso de representación conceptual.

3.1.1 Identificación de conceptos

La identificación de los conceptos que aparecen en un texto conlleva la realización de dos tareas principales:

- Identificar cada una de las palabras relevantes en el texto, determinando su categoría gramatical y su raíz o lema. Mediante el uso de un etiquetador léxico se determinan las palabras más relevantes (nombres, adjetivos, verbos y adverbios) que conforman el texto. Este etiquetador léxico (Acebo et al., 1994; Pla, 2000) proporciona además de la categoría gramatical, su raíz o lema.

- Establecer el significado correcto de cada una de estas palabras mediante un sistema de WSD que permita además obtener la información semántica asociada. La palabras polisémicas presentan el problema de la identificación del sentido correcto al que se hace referencia en el texto. Para resolver este problema se ha utilizado el sistema desarrollado por la Universidad de Alicante (Montoyo and Suárez, 2001; Suárez and Palomar, 2002; Montoyo, Palomar, and Rigau, 2001), sistema híbrido que combina los métodos de marcas de especificidad y máxima entropía. Una vez obtenido el sentido correcto se obtendrá de una

base de datos léxica, como EuroWordNet (Vossen, 1998), toda su información semántica asociada.

Estos pasos permiten definir la lista de los conceptos más relevantes incluidos en un texto, los cuales se representan gráficamente tal y como se muestra en la figura 1. Como se presenta en la sección 3.1.3, tal representación gráfica incluye tanto el significado correcto de la palabra como su información semántica asociada.

3.1.2 Identificación de relaciones y roles

Una vez definida la lista de conceptos, se deberá determinar las relaciones existentes entre ellos. Las únicas relaciones que se van a considerar en nuestra propuesta son:

- Que un concepto corresponda al rol conceptual de objeto.
- Que un concepto describa o sea propiedad de otro. Se identificará con el rol conceptual de atributo de objeto.
- Que un concepto mantenga una determinada relación con otro. Se identificará con el rol conceptual de relación.
- Que un concepto describa o sea propiedad de una relación entre conceptos. Se identificará con el rol conceptual de atributo de relación.

Como se presenta en (Burg and van de Riet, 1996), la tarea de identificar este tipo de roles ha pasado por diferentes etapas, como son el uso de heurísticas, el análisis gramatical o el análisis semántico. Actualmente el mecanismo más extendido para la detección e identificación de este tipo de relaciones entre conceptos es una combinación de estos métodos.

Sin embargo, teniendo en cuenta el proceso anterior de identificación de conceptos en el que se hace uso de información semántica y partiendo del análisis gramatical realizado, en esta fase sólo será necesario añadir el análisis sintáctico y la resolución de la ambigüedad referencial junto al uso de heurísticas del tipo *un nombre es un objeto* o *un adjetivo es una propiedad del objeto*, a fin de obtener un conjunto válido de patrones conceptuales. En este trabajo se ha utilizado la herramienta SUPPAR (Ferrández et al., 1999) que realiza al mismo tiempo el análisis sintáctico del texto junto a la resolución de la ambigüedad referencial. Los patrones conceptuales considerados en este trabajo son los siguientes:

1. un nombre común es una entidad u objeto.
2. si el verbo es conocer, mantener o saber[2], el objeto directo, sea un nombre común o adjetivo, representará a un atributo de la entidad.
3. cualquier otro verbo indicará una relación entre entidades
4. si el verbo es conocer, mantener o saber, y el objeto directo es ya una entidad, el verbo indicará una relación entre entidades.

La aplicación de estos patrones debe realizarse en el orden presentado tal y como se muestra en el siguiente apartado.

3.1.3 Representación conceptual

Una vez realizado el proceso de identificación de los conceptos y de las relaciones entre ellos, la representación conceptual de esta información se realiza a dos niveles. En el primer nivel se reflejan las relaciones entre conceptos que aparecen en el texto, como se muestra en la figura 3. En esta figura se puede observar que los objetos o entidades se representan como rectángulos de línea discontinua y las propiedades y relaciones como flechas de línea continua. Además, para el caso de relaciones estas flechas se etiquetan con el nombre de la relación.

En el segundo nivel se expande con la información que representa cada concepto tal y como se muestra en la figura 1. Esta figura muestra el concepto planta con el sentido relativo a una planta vegetal, tal y como se obtendría de completar el proceso anterior. En este nivel, las entidades u objetos siguen representándose mediante rectángulos de línea discontinua. Además, se añade información relativa a los sinónimos, representados con una línea continua, los hiperónimos, representados con una flecha de línea continua, y los merónimos con una flecha de línea discontinua.

Mediante esta representación conceptual se pretende realizar una representación

[2]La selección de estos verbos se debe al uso de textos orientados al diseño conceptual, los cuales presentan una estructura y vocabulario específicos.

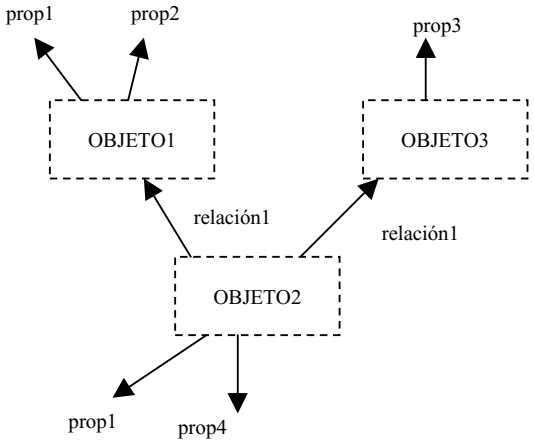

Figura 3: *Representación conceptual de un texto*

gráfica que facilite la compresión de un texto. Por ello, proponemos hacer dicha representación en dos niveles. En el primer nivel se representan exclusivamente los conceptos del texto con sus relaciones y en un segundo nivel se añade toda su información semántica, la cual ha sido adquirida mediante la aplicación de técnicas lingüísticas tal y como se ha explicado anteriormente.

3.2 Modelización conceptual de SI a partir de la representación conceptual propuesta

A continuación se presenta un pequeño caso de estudio sobre el que se aplicará la metodología descrita a la modelización conceptual de un sistema de información. Considérese el siguiente texto:

"Se desea mantener información relativa a las plagas que atacan, o han atacado, a los diferentes países de la Comunidad Europea. Para ello, de las plagas se desea conocer, su nombre, resistencia, organismos a los que afecta, los diferentes tipos de daños que puedan causar y los plaguicidas que la combaten. Con respecto a los organismos, se desea saber el nombre y el tipo genérico al que pertenece. Es importante destacar que un mismo organismo puede verse afectado por diferentes plagas. De los plaguicidas, se desea conocer su nombre, modo de acción, constitución química y otras características que se consideren relevantes. Puede darse el caso de que un mismo plaguicida combata muchas plagas, pero cada una con un grado de eficiencia diferente."

La aplicación del etiquetador léxico determina la lista de las palabras relevantes (nombres, adjetivos, verbos y adverbios). WSD nos proporciona el sentido de cada una de las palabras polisémicas de la lista anterior. Se debe resolver el fenómeno lingüístico de la ambigüedad referencial, como por ejemplo en la frase: "*Con respecto a los organismos se desea conocer el nombre y el tipo genérico al que pertenecen*" se obtendrán las frases: "*Con respecto a los organismos se desea conocer el nombre*" y "*Con respecto a los organismos se desea conocer el tipo genérico al que pertenecen*" lo que da lugar a las dos listas de conceptos siguientes: "*organismo conocer nombre*" y "*organismo conocer tipo*".

Mediante la aplicación de los patrones conceptuales vistos en la sección anterior se llegará a la conclusión de que la entidad organismo tiene como atributo la propiedad nombre y la propiedad tipo.

Considérese, ahora, la frase "*Las plagas afectan a organismos*". Si a dicha frase se le aplican las mismas técnicas que las aplicadas en las frases anteriores, se obtendrá la lista de conceptos "*plaga afectar organismo*'. Y mediante la aplicación de los patrones conceptuales se llegará a la conclusión de que la entidad plaga se relaciona con la entidad organismos, indicando con ello qué plagas afectan a qué organismos.

Si se realiza el mismo proceso con el resto de frases del texto se llega a la conclusión de que en dicho texto se pueden identificar varios objetos, como, plaga, organismo, daño y plaguicida, cada uno de los cuales se describen por una serie de propiedades:

- plaga: nombre, resistencia
- organismo: nombre, tipo

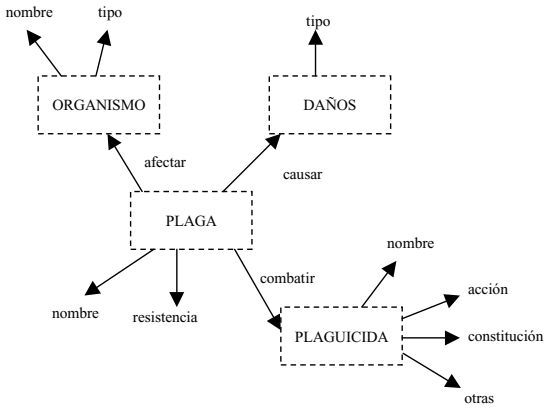

Figura 4: *Representación conceptual del ejemplo*

- daños: tipo
- plaguicida: nombre, modo de acción, constitución química y otras características

También, se pueden identificar varias relaciones entre objetos, como son:

- Afectar: un organismo puede verse afectado por varias plagas
- Causar: una plaga puede causar muchos tipos de daños
- Combatir: una plaga puede ser combatida por diferentes plaguicidas, cada uno de ellos con un grado de eficiencia diferente

Cada uno de estos elementos básicos indicados (objetos, propiedades y relaciones) serán representados gráficamente conforme al criterio preestablecido anteriormente. La figura 4 corresponde a la información relativa al primer nivel.

4 Aplicaciones inmediatas

La aplicación más inmediata de este proceso es la generación automática de modelos de datos a partir de textos. Para ello, bastaría con modificar la forma de representación de los concepto a la metodología deseada.

Otra de las aplicaciones inmediatas sería la recuperación de información. Se trata de crear una representación conceptual tanto de la pregunta como de los documentos, que permitirá compararlos no sólo usando una medida de co-ocurrencia sino también a través de la similitud conceptual de la pregunta y los documentos. Se puede realizar una búsqueda inicial utilizando la representación de primer nivel y posteriormente este resultado se filtra con la información semántica asociada que aparece en el segundo nivel. Este es el enfoque que se sigue en el proyecto "TUSIR: Desarrollo de un sistema de comprensión de textos aplicado a la recuperación de información"`http://gplsi.dlsi.ua.es/TUSIR/`

5 Conclusiones y Trabajos futuros

En este artículo se ha presentado una propuesta de representación conceptual utilizando técnicas lingüísticas tales como el etiquetado léxico, la desambiguación del sentido de las palabras, la resolución de la ambigüedad referencial y el análisis sintáctico. Dichas técnicas permiten representar la información contenida en un texto de forma estructurada.

El mecanismo de representación automático aquí planteado, presenta una serie de limitaciones como son:

- Sólo se permiten o se consideran relaciones entre pares de objetos.
- No se ha considerado información relativa a la cardinalidad de las relaciones. Es decir, se han ignorado frases como "*Puede darse el caso de que un mismo plaguicida combata muchas plagas.*"
- Los textos considerados, son textos orientados al diseño conceptual, por lo que presentan una estructura y vocabulario específicos.

Todo ello abre un gran marco de trabajos futuros que determinen:

- Cómo detectar y representar clasificaciones de objetos haciendo uso de los recursos semánticos, tales como ontologías.
- Cómo detectar y representar relaciones entre un número variable de objetos.
- Cómo establecer y representar la cardinalidad de las relaciones.

Además, habría que contemplar y analizar más variedad de textos que permitieran la definición de nuevos patrones conceptuales, independientes de la estructura de dichos textos.

Bibliografía

Acebo, S., A. Ageno, S. Climent, J. Farreres, L. Padró, F. Ribas, H. Rodríguez, and O. Soler. 1994. Maco: Morphological analyser corpus oriented. Acquilex II WP 31.

Burg, J.F.M. and R.P. van de Riet. 1996. Analyzing Informal Requirements Specifications: A First Step towards Conceptual Modeling. In *Procedin of Second International Workshop on Applications of Natural Language to Information Systems (NLDB'96)*, Amsterdam, The Netherlands.

Ferrández, A., M. Palomar, P. Martínez-Barco, J. Peral, R. Muñoz, and M. Saiz-Noeda. 1999. Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística. *Procesamiento del Lenguaje Natural*, 25:217–218.

Montoyo, A., M. Palomar, and G. Rigau. 2001. Method and Interface for WordNet Enrichment with Classification Systems. *Lecture Notes in Computer Science*, 2113.

Montoyo, A. and A. Suárez. 2001. The University of Alicante word sense disambiguation system. In Judita Preiss and David Yarowsky, editors, *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 131–134, Toulouse, France, July. ACL-SIGLEX.

Muñoz, R. and M. Palomar. 2001. Clustering technique based on semantic for definite description resolution. In *Proceedings of the International Conference Text Speech and Dialogue*, volume 2166 of *Lecture Notes in Artificial Intelligence*, Czech Republic. Springer-Verlag.

Overmyer, S.P., B. Lavoie, and O. Rambow. 2001. Conceptual Modeling through Linguistic Analysis Using LIDA. In *ICSE*. IEEE Computer Society.

Palomar, M., A. Ferrández, L. Moreno, P. Martínez-Barco, J. Peral, M. Saiz-Noeda, and R. Muñoz. 2001. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics*, 27(4):545–567.

Palomar, M., A. Ferrández, L. Moreno, M. Saiz-Noeda, R. Muñoz, P. Martínez-Barco, J. Peral, and B.Ñavarro. 1999. A Robust Partial Parsing Strategy based on the Slot Unification Grammars. In *Proceeding of 6e Conférence annuelle sur le Traitement Automatique des Langues Naturelles. (TALN'99)*, pages 263–272, Cargèse, Corse.

Pla, F. 2000. *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos.* Tesis doctoral, Departamento de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia.

Suárez, A. and M. Palomar. 2002. A maximum entropy-based word sense disambiguation system. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, August. (to be published).

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, 3(1).