Extracción Automática de Relaciones Semánticas

Sonia SÁNCHEZ-CUADRADO,
Juan LLORÉNS,
Jorge MORATO
y
José A. HURTADO
Departamento de Informática, Universidad Carlos III
Leganés (Madrid) 28911, España

RESUMEN

Se presenta una metodología para la creación automática de ontologías de tipo tesauro. La información es extraída mediante procesamiento del lenguaje natural de un corpus documental de un dominio. De este modo, el presente estudio se ha centrado en interrelacionar, de forma automática, los conceptos designados por los términos de los documentos, dependiendo de las relaciones semánticas que se establecen entre ellos. En el presente documento se muestra los resultados de la aplicación del método al dominio de la zoología. Se resumen estadísticamente los logros alcanzados, reseñando los problemas encontrados y la identificación de aquellas estructuras que aportan resultados más fiables.

Palabras Claves: Procesamiento del Lenguaje Natural, Taxonomías, Creación Automática de Ontologías, Relaciones Semánticas

INTRODUCCIÓN

Es una realidad que el volumen de información ha aumentado en una proporción desmesurada en los últimos tiempos. No obstante, un proceso más espectacular lo ha protagonizado la incorporación masiva de usuarios con multitud de necesidades diferentes de información, sobre todo en lo referente a Internet. Estos fenómenos tienen dos requisitos fundamentales, la Organización de la Información (*Information Organization*) y la Organización del Conocimiento (*Knowledge Organization*).

Algunas de las propuestas orientadas hacía la Organización de la Información y la Organización de Conocimiento están basadas en recursos como tesauros, taxonomías y ontologías. Estas suponen un recurso idóneo para el almacenamiento, la comunicación y la recuperación de información del conocimiento especializado de una disciplina [1], [2], [3].

Construcción automática de Tesauros y Ontologías

Tesauros y Ontologías han sido creados tradicionalmente de forma manual. Sin embargo, las iniciativas actuales tratan de construir estos recursos de forma automática, por medio de la selección de la terminología y de las relaciones que se establecen entre los conceptos del dominio. Entre las iniciativas para la construcción automática de ontologías destacan *Lexical FreeNet*[1] como ejemplo de ontología construida de forma automática, y *EWN* o *Sensus*[2] por su enfoque semiautomático. Existen además distintas aplicaciones para ayudar a la construcción de este tipo de recursos, como es el caso de Text-to-Onto para las ontologías [4] o SEXTANT para los tesauros. SEXTANT está basada en el trabajo de Grefenstette [5]. En este trabajo, se asignan pesos, dependiendo de la frecuencia con la que se encuentra una relación. Grefenstette sugiere, como última etapa del proceso, una comparación del resultado con otros recursos elaborados manualmente.

El análisis de las tecnologías aplicadas para la construcción de tesauros y ontologías de forma automática muestra que se emplean modelos mixtos de Procesamiento del Lenguaje Natural (PLN) y teorías matemáticas de la comunicación aplicados al análisis automático de contenido. La obtención relevante de información de un texto especializado se realiza mediante la selección de estructuras y datos implícitos o explícitos dentro del campo de la Extracción de Información (*Information Extraction*) y la Semántica.

Identificación y clasificación semántica de estructuras fraseológicas

Para el inglés, existen diversos estudios dedicados a identificar y clasificar las relaciones que se establecen entre los conceptos. Algunos estudios, como los de Nakumara [6], estaban encaminados, inicialmente, a un proceso manual, pero cada vez más, por parte de otros autores (Green [7], Hearst [8]) surgen nuevas aportaciones orientadas a la semiautomatización o automatización de estos procesos.

Hearst [8] y Grefenstette [9] en sus estudios y experimentos encuentran relaciones -pares de hiperónimos e hipónimos- en los textos, usando patrones (*pattern-matching*). Consideran que estas estructuras léxico-sintácticas son fácilmente reconocibles y que ocurren con cierta frecuencia en los textos, proporcionando relaciones de interés.

Básicamente, la identificación de relaciones se ha centrado en dos aspectos, entre otros; las relaciones sintagmáticas y las relaciones paradigmáticas. Tanto las paradigmáticas como las

[1] Lexical FreeNet. http://www.lexfn.com [consultado Julio 2004]
[2] SENSUS http://www.isi.edu/natural-language/resouces/sensus.html [consultado Julio 2004]

sintagmáticas tienen diferentes métodos para tratar de identificarlas mediante axiomas de función y entidad.

En el caso de las relaciones sintagmáticas se ha generalizado el uso de verbos para su localización y extracción de propiedades. Otros autores como Oishi y Matsumoto [10] han propuesto un mapeo de alternancia de patrones a nivel superficial de subcategorización de marcos para la adquisición de roles temáticos. Rebecca Green [7] estudia los factores y las características que deben cumplir las relaciones sintagmáticas para poder relacionarse con ciertos roles temáticos.

También para el español se han realizado investigaciones con fines similares, por algunos autores, como Martì [11] y Díez [12]. Por otra parte, Lorente [13] ha comprobado que existen expresiones simples y complejas en español que permiten identificar el tipo de relación existente entre dos conceptos en un dominio. Los contextos en los que se expresan los conceptos son normalmente complejos y variados, e incluso constituyen múltiples formas simples interrelacionadas. Este planteamiento conlleva dificultades a la hora de identificar los conceptos y sus relaciones, hasta el punto de que algunas de las primeras investigaciones se aplicaron para tipos de textos muy estructurados, tales como los diccionarios en inglés [14], [15] y en español [11].

Estos autores coinciden en afirmar que este enfoque supone un avance hacia la Organización del Conocimiento. Sin embargo, ni la identificación de las expresiones relacionales en LN, ni la asignación del tipo de relación semántica supone una tarea sencilla. En la actualidad, existen propuestas diferentes para las clasificaciones de las relaciones semánticas [1], [16], [17]. Estas relaciones dependen de las características léxico-semánticas y sintáctico-semánticas que se establecen entre los conceptos o del fin que se persiga con respecto al dominio tratado.

METODOLOGÍA

Con este trabajo se pretende presentar y analizar un método de extracción automática de relaciones semánticas sin necesidad de documentos con una rigurosa estructura. La metodología se ha centrado en determinar los métodos automáticos de extracción, una vez seleccionadas la semántica de las relaciones a analizar, y sus correspondientes unidades fraseológicas.

El sistema automático consta de un proceso de indización capaz de reconocer relaciones sintagmáticas. Se realiza mediante un análisis superficial de la frase con estructuras sintáctico-semánticas apoyado por roles de las preposiciones y con patrones asociados de las estructuras reiterativas de los textos. Dada la ambigüedad de la lengua y de las formas de expresión de las relaciones, se han estudiado características de las estructuras relacionales capaces de establecer vínculos semánticos más fiables.

Los pasos seguidos para lograrlo son:

- Asignar tipos de relaciones semánticas a estructuras léxicas verbales que identifiquen los roles de los sustantivos que la acompañan. Para la identificación de estas estructuras se ha recurrido a revisiones de la literatura sobre el tema [7], [18], y a la inclusión de nuevas relaciones ad-hoc, basadas en el análisis de las unidades fraseológicas de los textos.

- Para evaluar las relaciones extraídas se ha recopilado una colección de documentos (corpus documental). Se ha creado una aplicación cuyo fin es indizar y extraer automáticamente las relaciones semánticas del documento. El objetivo es comprobar qué relaciones semánticas son identificadas en el corpus por estar presentes en una estructura sintagmática.

Asignación de tipos de relaciones a estructuras léxicas verbales

El planteamiento de identificar estructuras léxicas verbales de carácter relacional tiene como objetivo la creación de un corpus de estructuras que se puedan asociar con un rol semántico. Para el módulo de Creación del Corpus de Estructuras Semánticas Relacionales (CESER) se ha partido de los estudios de la literatura sobre el tema para el inglés [7] y para el español [11], [12], [13]. Según estos estudios la tipología de las relaciones semánticas identificadas dependerá de varios factores:

- **Dominio**: aunque las unidades fraseológicas tienen carácter genérico, es decir son independientes de un determinado dominio de conocimiento, las investigaciones reseñadas apuntan a que existen estructuras que pertenecen a dominios concretos

- **Grado de estructuración del documento**: un documento que tenga una estructura rígida permite identificar más claramente las relaciones, mientras que el discurso libre supone una dificultad añadida, por no poseer más rigidez de la que permite la sintaxis del español.

- **Tipología Documental**: puede influir en los patrones que se establezcan para su posterior extracción automática.

- **Finalidad del análisis del domino**: puede hacer variar la prioridad de las relaciones semánticas a identificar.

- **Estilo del autor:** la forma de expresarse del autor influye en las presencia en los documentos de ciertas unidades fraseológicas.

Tras un estudio de la literatura sobre el tema, así como un muestreo de las unidades fraseológicas presentes en distintos documentos y su posible correspondencia semántica, se han propuesto para este trabajo las siguientes relaciones:

- Jerarquía o generalización
- Equivalencia
- Todo parte o meronimia
- Entidad-localización
- Autoría-entidad
- Entidad-proceso

Una vez seleccionadas las relaciones semánticas y las estructuras que permiten localizarlas, se han establecido manualmente los patrones que permiten su identificación. Para ello, no sólo se hace necesario identificar manualmente estas correspondencias entre determinados patrones e información semántica, sino que se necesita implementar un analizador del Lenguaje Natural (Figura 1).

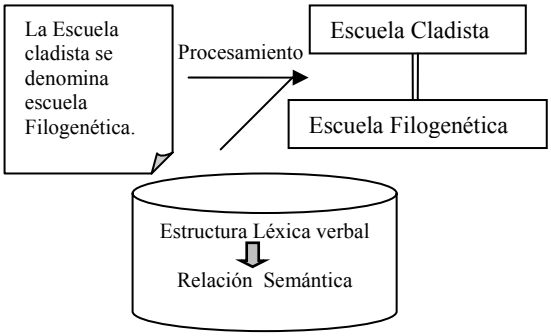

Figura 1. Ejemplo de aplicación del recurso desarrollado

Para la implementación del analizador en Lenguaje Natural se han tenido en cuenta elementos como:

- Verbos y estructuras léxicas verbales. En las estructuras léxicas verbales los verbos pueden estar asociados a sustantivos y preposiciones. Las estructuras léxicas verbales tienen asignada la semántica que relaciona los elementos de la frase. Por ejemplo, "estar formado por", "estar compuesto por" y "formar parte de" pueden dar una relación de meronimia.

- Preposiciones. Se han definido varios subtipos de preposiciones según su funcionalidad para enlazar o relacionar los distintos complementos de la frase.

- Negaciones. Se consideran las negaciones por cuanto pueden condicionar el carácter de determinada relación o incluso variarla.

- Elementos de probabilidad (adverbios como frecuentemente, usualmente o verbos como soler, poder, etc). Estos elementos tienen gran importancia, ya que modifican la probabilidad de que una relación sea relevante.

- Además se han tabulado otros elementos como: locuciones, adverbios o adjetivos.

La presencia de determinados signos de puntuación: p.e. comas o dos puntos en enumeraciones, etc., se ha tenido en cuenta a la hora de relacionar los elementos de la frase.

El módulo de Creación del Corpus de Estructuras Semánticas Relacionales (CESER) está integrado dentro de una herramienta de gestión y análisis de dominios.

Evaluación de la correlación estructura fraseológica-relación semántica

La herramienta de gestión y análisis de dominios creada para recoger las relaciones de los documentos, sigue las siguientes etapas:

1. Creación de una colección de documentos sobre el dominio a estudiar.

2. Identificación del vocabulario:

Equivalencia

   a. Lematización de cada una de las palabras procesadas.

   b. Extracción del tipo gramatical correspondiente de cada palabra.

   c. Creación e identificación de palabras compuestas.

3. Filtrado del vocabulario: es decir, selección de los términos más representativos del dominio entre el total de términos analizados.

4. Identificación de las relaciones entre los términos del vocabulario, basada en la correlación entre estructuras fraseológicas y relaciones semánticas descrita previamente.

5. Cálculo de asignación de pesos como ayuda para el filtrado y validación de relaciones.

Selección del corpus: Se han creado dos corpus o colecciones de documentos para realizar el experimento:

- Corpus del dominio: Para evaluar la metodología, se ha creado un corpus documental con documentos de carácter científico-técnico. El corpus está compuesto de treinta documentos pertenecientes al dominio de la zoología sistemática. Este dominio ha sido seleccionado debido a su predecible riqueza en relaciones de tipo tesauro (jerarquía, meronimias y sinonimias principalmente), además de poseer un vocabulario altamente normalizado.

- Corpus de comparación: Se trata de una colección de documentos de carácter multidisciplinar. Su función es detectar los términos del corpus del dominio que tienen una frecuencia significativamente diferente a la esperada en un dominio multidisciplinar.

Extracción automática de palabras simples: Mediante el sistema informático de tratamiento de la información textual y gráfica producida por las aplicaciones ofimáticas mas comunes, se consigue extraer las palabras simples que forman los documentos del corpus.

Lematización: El objeto de este módulo es normalizar las palabras de un documento de texto en Lenguaje Natural a una forma canónica. El módulo sigue los siguientes pasos:

- Generar todas las variantes flexionadas de cada término de entrada en función de un conjunto de reglas de lematización para el español.
- Confrontar cada variante flexionada contra un vocabulario del mismo idioma con vistas a identificar su categoría morfológica real.
- Proceso de normalización: las distintas variantes de palabras flexionadas y derivadas son unificadas bajo una forma normalizada, que permite eliminar las distintas variantes debidas al número, género, etc.
- Por lo tanto, las estructuras léxicas verbales de carácter relacional quedarán también normalizadas a una forma común. En un primer momento, las formas verbales normalizadas al infinitivo corresponden con las formas personales del verbo. Las formas no personales del verbo en especial el participio son tratadas de forma diferente dado que entendemos que reflejan con más validez las relaciones semánticas.

Identificación de términos compuestos: El uso de un vocabulario controlado basado en unitérminos supone a menudo un aumento de la ambigüedad terminológica que la creación del tesauro pretende evitar. Con esta finalidad este módulo construye términos compuestos que son por lo tanto más específicos. Para ello se tiene en cuenta la identificación de las categorías gramaticales de los elementos implicados en la construcción del término compuesto y de aquellos elementos susceptibles de ser agrupados.

La principal ventaja de esta fase reside en dos cuestiones. Por una parte poder encontrar conceptos representados por términos compuestos, y en segundo lugar poder encontrar relaciones de forma automática entre los términos compuestos, y sus términos simples constituyentes. Esta relación estructural –no basada en verbos-, se obtiene al identificar como elemento genérico de "taxonomía animal" el término "taxonomía". Esta última cuestión sólo en caso de que los conceptos simples se consideren como relevantes para el dominio.

Filtrado Terminológico: Se basa en las siguientes estrategias:

- Filtrado basado en TF-IDF para identificar y validar manualmente aquellos términos con escasas posibilidades de pertenecer al dominio.
- Módulo de filtrado asistido. Este módulo permite tratar de forma masiva con términos especialmente problemáticos. Este módulo muestra cada uno de los términos con n caracteres, permitiendo mediante un gestor de terminología eliminarlos, editarlos o unificarlos con otro término del vocabulario.
- Comparación de los valores de tf e IDF contrastándolo con el corpus de comparación.

Este apartado se subdivide en las siguientes tareas:

- Cálculo de la frecuencia de cada término normalizado en distintos documentos del corpus del dominio.
- Comparación del dato anterior con el que sería esperable en un corpus de comparación. El corpus de comparación se utiliza para comparar la frecuencia de los términos en esta colección, con la obtenida en el procesamiento del corpus documental y como base para localizar unidades léxicas. Cuando la frecuencia relativa es similar, se presupone que el término pertenece al vocabulario científico-técnico del idioma y no al vocabulario específico del dominio de estudio.

Correlación entre Sintagmas Verbal y Relación Semántica: El objetivo de esta fase pretende ser la identificación automática de relaciones empleando las relaciones previamente establecidas entre estructuras semánticas y sintácticas.

La correspondencia entre estructuras fraseológicas y relaciones semánticas no es siempre uno a uno, sino frecuentemente muchos a muchos. Es decir, de una misma sentencia se pueden suponer varias posibles relaciones semánticas entre los complementos, por lo que es necesario un sistema que ayude a discriminar cuál es la relación más plausible. Con este fin se ha creado un sistema que asigna pesos a las relaciones, ayudando al experto en el dominio (también denominado ingeniero de dominio) en la toma de decisiones. El sistema para el cálculo de pesos se desarrolla en el siguiente apartado.

Filtrado relacional y problemas en la identificación de relaciones: Una vez identificado el vocabulario básico debe hacerse, como en la fase de extracción terminológica, un filtrado de las relaciones. Se han identificado tres grupos de errores que deberán corregirse manualmente:

1. Relaciones con identificación errónea de la relación semántica entre términos. El sistema alerta del error pero debe ser depurado manualmente. Un ejemplo: "Juan nació en Francia" y "Juan nació en México" no son relaciones contradictorias en el caso de tratarse de diferentes instancias de Juan. También, podrían deberse a contradicciones en los documentos, bien sean de redacción o por información equivocada. En cualquier caso la asignación de pesos puede ayudar al experto en el dominio a discriminar aquella que es correcta.
2. Relaciones contradictorias por su naturaleza. En algunos campos como las opiniones políticas o las escuelas de pensamiento se pueden extraer relaciones contradictorias que no supongan un error radicado ni en el material de partida ni en la aplicación.

3. Relaciones implícitas y explicitas. Pueden existir relaciones que no aparezcan en los documentos por considerarse información obvia. Por ejemplo es relativamente complicado en documentos que no sean obras de referencia encontrar la relación "perro" <es un tipo de> "animal". Esta información deben introducirse manualmente o ampliar el corpus con documentos que puedan tenerlas (p.e. diccionarios).

4. Relaciones detectadas como erróneas por suponer ciclos del tipo: "A genérico de B" y "B genérico de A", pueden deberse a errores de la aplicación, de los documentos o a problemas de cardinalidad múltiple comentados más arriba entre de unidades fraseológicas y relación semántica. Un caso similar se da cuando se identifican dos relaciones distintas entre dos términos: "A genérico de B" y "A sinónimo de B". En ambos casos deben de ser corregidos a mano, si bien se puede utilizar como elemento decisorio la asignación de pesos explicada a continuación.

5. Relaciones que no aportan información relevante sobre el dominio o que poseen menor fiabilidad. También pueden ser identificadas por el proceso de asignación de pesos.

Para detectar estos problemas y como herramienta para corregirlos, se ha diseñado un módulo que permite visualizarlas y gestionarlas (Figuras 1 a 3). Además se ha reformulado tf-IDF aplicado a relaciones.

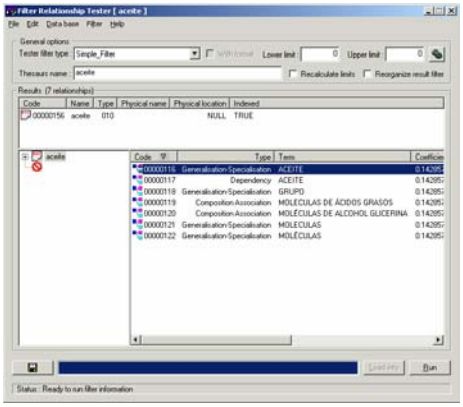

Figura 2. Sistema de filtrado y modificación de relaciones

Se han tenido en cuenta tres parámetros para asignar pesos a las relaciones:

- IDFR (IDF Relacional) se calcula como el número de documentos en la colección dividido por número de documentos con la relación i en la colección.
- Fr (Frecuencia de la Relación) como el número de veces que aparece determinada relación en los documentos de la colección.
- Vr (Estructura verbal de la Relación) como el número de estructuras fraseológicas verbales diferentes que indican un mismo tipo de estructura verbal entre dos términos dados.

De este modo, será más fiable una relación que aparece en varias estructuras verbales relacionales diferentes que indican la misma relación que la que aparece en un único documento o bajo una única estructura verbal.

La existencia de un elemento de probabilidad (p.e. usualmente, frecuentemente, raramente, etc.) puede atenuar o aumentar el peso de una relación.

En el caso de que el estudio de la frecuencia no sea lo suficientemente relevante, existe la posibilidad de modelar el sistema para que alguno de los términos asuma un peso superior por su formato, bien sea la tipografía (negrilla, subrayado, etc.) o, bien sea su localización en el documento original (título, resumen, etc.).

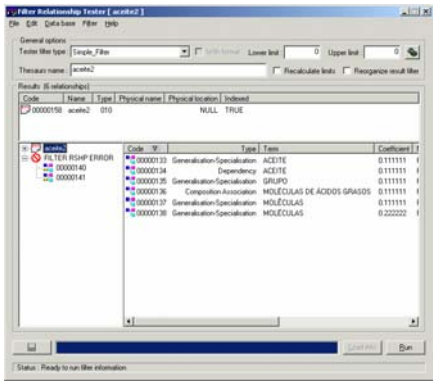

Figura 3. Sistemas de detección de relaciones erróneas

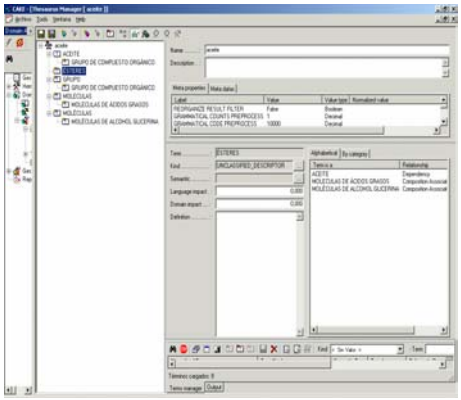

Figura 4. Sistema de Organización de Relaciones

En la Figura 4 se muestra un ejemplo del resultado organizado con la estructura de tesauro de la metodología señalada, permitiendo así revisar los datos y completar las relaciones entre los descriptores.

RESULTADOS

Se ha realizado una extracción automática de las estructuras relacionales indicadas más arriba en el dominio de la zoología sistemática. A continuación se han tipificado manualmente relacionando las estructuras sintácticas con roles semánticos entre términos.

Proceso de filtrado mediante comparación de frecuencias entre corpus del dominio y corpus de comparación

Mediante comparación con una colección de un dominio general se ha obtenido las siguientes medidas:

- Términos obtenidos en la colección de comparación: 21.298, el número tan elevado de términos se debe sin duda por pertenecer a un dominio multidisciplinar.
- Términos obtenidos en el corpus de zoología: 8.109
- Términos comunes entre colección de comparación y el dominio estudiado: 1.412. Estos términos, suponiendo ningún solapamiento entre el corpus del dominio y el de comparación deben revisarse por su posible pertinencia a un vocabulario no específico. En el presente estudio, esta depuración se hizo manualmente.
- Términos con una frecuencia significativamente menor en la colección de comparación que en zoología: 932. Estos términos han sido filtrados del vocabulario obtenido en el corpus de zoología.

Relaciones

El sistema identificó 2.104 estructuras léxicas verbales de carácter relacional. De las cuales 84 tenían una semántica previamente asignada. Estas 84 estructuras se pueden agrupar en 32 subcategorías definidas a partir de las relaciones indicadas más arriba. La siguiente tabla muestra las relaciones más frecuentes:

Categoría	Estructuras con la categoría	Frecuencias localizados en el texto
Equivalencia	7	163
Generalización	5	142
Todo-parte/meronimia	6	118
Entidad-localización	11	96
Autoría-entidad	4	92
Entidad-proceso	1	61

Tabla 1. Número de estructuras fraseológicas mapeadas por cada relaciones semánticas

Estudio de equivalencias

Se ha realizado un seguimiento de las relaciones de equivalencias, por su implicación en la creación de tesauros y ontologías donde se encuadra este estudio. En primer lugar se han marcado una serie de estructuras léxico verbales que podrían establecer una relación de sinonimia. A continuación se ponen algunas de estas estructuras de forma normalizada, esto significa sin palabras identificadas como vacías y en una forma verbal unificada a infinitivos, participios y formas pronominales:

- Sinónimos absolutos: denominado/s, ser un equivalente de, ser igual que, ser equivalente a, ser un equivalente de, ser lo mismo que, ser idéntico a, equivaler a, ser sinónimo de, ser un sinónimo de, llamado/s, recibir el nombre de,...
- Relaciones de equivalencia: estar relacionar, simbolizar mismo, indicar, denotar mismo, semejar, significar, significar mismo, querer decir, ser similar, ser parecer, asemejarse, ser análogo, indicar el/lo mismo, denotar, traducir, significar, conocer como, considerar a ... como, poder considerarse, poder ser considerado como, ir a considerar, ser traducibles a, que tratar de unificar, ...

Hay que tener en cuenta que, debido a la correspondencia de varios a varios entre frases y semántica, algunas de las relaciones semánticas verbales propuestas más arriba pueden también dar lugar a otras relaciones semánticas.

A continuación, se han buscado automáticamente estas estructuras en el texto. Una proporción muy pequeña de las estructuras creadas se han identificado, en concreto las siguientes:

Estructura Relacional	Ocurrencias
Llamado	84
Denominar	29
Denominado	29
Conocer como	9
Conocido	7
Conocido como	4
Recibir el nombre	1
Haber denominado	1

Tabla 2. Estructuras verbales más frecuentes que han sido mapeadas contra relaciones semánticas

Es decir, muchas de las estructuras más significativas no han sido localizadas en el texto, este hecho puede deberse, entre otros, al estilo del autor o al dominio tratado.

Por otra parte, los resultados al validar la pertinencia de las relaciones son:

- Un ejemplo significativo en el análisis de las estructuras léxico verbales de carácter relacional, es el participio "llamado". Indica en los 84 casos estudiados una relación de equivalencia. De los 84 casos, 74 fueron correctamente identificadas por el módulo automático. Frecuentemente se observa que se puede inferir una relación genérico-específico a partir de la de equivalencia. p.e. a partir de

"*... cordón sinuoso llamado filamento mesentérico ...*"

se puede relacionar que un tipo de"cordón" es el "filamento mesentérico" ya sea por la relación de jerarquía entre "cordón" y "cordón mesentérico" y este con la equivalencia entre "cordón mesentérico" y "filamento mesentérico".

- Con algunos participios, como es el caso de "denominado" de las 29 apariciones, el sistema no pudo extraer la relación de seis, debido a:
    - anáforas
    - errores tipográficos
    - estructuras no identificadas por el sistema por no haberse localizado en la colección de comparación o por generar situaciones ambiguas

CONCLUSIONES

Se ha observado que la metodología diseñada funciona mejor en dominios específicos donde el nivel de ambigüedad terminológica y relacional es inferior que en dominios genéricos.

La elección del corpus tiene un papel clave en el funcionamiento de la metodología, por cuanto determinados dominios como la zoología sistemática proporcionan términos más específicos o nada frecuentes en otros dominios. Este hecho resulta útil para realizar un filtrado mediante una colección de comparación.

Se han detectado tres tipos de unidades léxicas verbales:

- Aquellas que establecen correspondencia entre dos conceptos;
- Aquellas que aportan contenido temático de la especialidad;
- Verbos que trasmiten información referida a los contenidos de determinado dominio.

Algunos ejemplos de las relaciones semánticas identificadas entre términos son la sinonimia, hiponimia, hiperonimia, meronimia, y tipos diferentes de asociación, tales como localización, agente, causa, instrumento, etc., que proporcionan información relevante para una posterior aplicación a los sistemas de recuperación.

Se ha observado que las relaciones obtenidas con formas no personales de verbos parecen aportar relaciones más fiables que las formas personales. En futuros desarrollos de la herramienta se pretende incorporar este factor en la asignación de pesos a las relaciones.

Por último, se ha verificado la conveniencia de que expertos en el dominio a estudiar supervisen distintas etapas, como: selección del corpus del dominio y la validación, adición y eliminación de términos y relaciones.

AGRADECIMIENTOS

Este trabajo se enmarca dentro del proyecto de investigación y desarrollo tecnológico titulado "Desarrollo de un tesauro de verbos para entornos de información dinámica. Aplicación del estándar ISO/IEC: 13250:1999". Está financiado por el Plan Nacional de I+D+I 2000-2003 (TIC2000-0383).

REFERENCIAS

[1] ISO. **Guidelines for the establishment and development of monolingual thesauri: international standard ISO 2788**, 2nd ed. 1986-11-15. [Geneve]: ISO, 1986.

[2] Gruber T. R., "Ontolingua: A mechanism to Support Portable Ontologies". **Report KSL 91-61**, Stanford University, 1992.

[3] Rodriguez, H., "Adquisición Automática y Uso de Taxonomías de amplia Cobertura". Tutorial presentado en **Workshop on Automatic Acquistion of Linguistic Knowledge**. San Millan, 2000. http://www.lsi.upc.es/~horacio/docencia.html [consultado 18/05/03]

[4] Gómez-Pérez, A, Fernández-López, M, Corcho, O, **Ontological Engineering**. London: Springer-Verlag, 2004.

[5] Grefenstette, Gregory, **Explorations in Automatic Thesaurus Discovery**. Boston: Kluwer Academic Publishers, 1994

[6] Nakumara, Y. "A Language for Knowledge Representation". **Advances in Knowledge Organization**. Vol.4, 1994, pp. 127-133.

[7] Green, R. "The Role of Relational Structures in Indexing for the Humanities". **Knowledge Organization**. *Vol.* 24 , No. 2, 1997, pp. 72-28.

[8] Hearst, M. "Automatic Acquisition of Hyponyms from Large Text Corpora". **Acts de COLING**, Nantes, 1992.

[9] Grefenstette, G. and Hearst, M.A, "A method for refining automatically-discovered lexical relations Combininig weak Techniques for Stronger Results". In **Proceedings of the Workshop on Statistically-Based Natural Language Programming Techniques**. Menlo Park, CA: AAAI Press, 1992

[10] Matsumoto & Utsuro, "Lexical knowledge Acquisition" pp. .563-610. in **Handbook of Natural Language Processing**. Ed: Dale, R. at. Marcel Dekker. New York, 2000

[11] Martí, A., Castellón, I., Fernández, A. "Extracción de información de corpus diccionariales". **Novática**, nº133 mayo-junio 1998, pp. 4-10.

[12] Díez, P. L. "La Relación de Meronimia en los Sustantivos del Léxico Español: Contribución a la Semántica Computacional". **Estudios de Lingüística Española**. Vol. 2, 1999.

[13] Lorente, M. "**Verbos y discurso especializado**". http://elies.rediris.es/elies16/Lorente.html. [consultado 01/07/04]

[14] Amsler, R. "A taxonomy for english verbs and nouns". In **Proceedings of the 19th Annual Meeting of the Association for Computacional Linguistics**. Stanford. CA., 1981, pp. 133-138.

[15] Copestake, Ane, "**Acquilex: The Acquisition of Lexical Knowledge**" www.cl.cam.ac.uk/Research/NL/acquilex/projdesc.html [consultado 22/05/04]

[16] Fillmore, C. J. "Types of lexical information". In: STEINBERG, D.D. and JAKOBOVITS, L.A., eds. **Semantics: and interdisciplinary reader in philosophy, linguistics and psychology**. Cambridge: Cambridge University Press, 1971.

[17] Tudhope, D; Alani, H. and Jones, C. "Augmenting Thesaurus Relationships: Possibilities for Retrieval". **Journal of Digital information**. Vol. 1, 2001.

[18] Sánchez-Cuadrado, S, Lloréns, J, Morato, J, "Desarrollo de una aplicación para la gestión de relaciones en tesauros generados automáticamente" **II Jornadas de Tratamiento y Recuperación de la Información (JOTRI'03)**. Madrid: Universidad Carlos III, 2003, pp. 151-156