Artículos

Utilización de estructuras verbales en la identificación de relaciones y descriptores en tesauros[1]

José Antonio Moreiro González
Juan Lloréns Morillo
Miguel Ángel Marzal García-Quismondo
Jorge Morato Lara
Sonia Sánchez Cuadrado
Pilar Beltrán Orenes

RESUMEN

Se parte del desarrollo que los tesauros automatizados han proporcionado a las llamadas relaciones no-clásicas de Van Slype, fomentando una eclosión de las relaciones de asociación, agrupadas en categorías por Tudhope, Alani, Jones y engrosadas por el traspaso de relaciones antes de jerarquía por la American Library Association (ALA). Frente a la estaticidad y clasificación taxonómica de los tesauros clásicos, inadaptados al dinamismo, transversalidad, exigencia de mayor abstracción de la información en hipermedia, se proponen las formas verbales como conceptos dinámicos, relacionales, conceptualizadores de la acción, muy útiles para representar el contexto mediante relaciones circunstanciales, a partir de los precedentes establecidos por A. L. Tharp y SYNTOL. Se indican como sus aplicaciones inmediatas la determinación de géneros de Swales, la tipología de las secciones del documento y unas funciones tesaurales más eficaces en red. Se concluye con futuros desarrollos de investigación en las categorizaciones asociativas, desde las experiencias aplicativas deWordNet.

ABSTRACT

The background of the work presented in this paper is based on automatic thesauri, and the improvement they provided to the "so called" not-classical relationships proposed by Van-Slype. We intend to enhance the usage of association relationships, grouped by Tudhope, Alani y Jones and emphasized by the American Library Association (ALA) relationships translation. We fight the static approach, as well as the taxonomical classification presented in classical thesauri, not allowing them to represent dynamics, transversal relationships, and hypermedia navigation. This paper propose "verbal forms" as dynamic concepts, origin to relationships, essentials of action, and very useful to represent context by circumstantial relationships (based on the antecedents of A. L. Tharp and SYNTOL). The main application areas of our work are gender determination of Swales, document sections typology and a more effective thesauri functions in the net. Future research developments include associative categorization from WordNet applicative experiences.

1 *Este trabajo ha sido realizado dentro del marco del Proyecto financiado por la CICYT (Comisión Interministerial de Ciencia y Tecnología), titulado "Desarrollo de un tesauro de verbos para entornos de información dinámica. Aplicación del estándar ISO/ICE: 13250:1999, del Plan General del Conocimiento. TIC 2000-2003.*

Introducción

Revisión de las relaciones tesaurales

La norma ISO 2788 (1986)[2] marca que las relaciones básicas de un tesauro son de tres tipos: equivalencia, jerárquica y asociativa. En el presente documento se hace una revisión de los principales cambios que han tenido lugar desde la publicación de la norma en las relaciones empleadas en los tesauros, para a continuación proponer un enfoque que emplea relaciones diferentes, atendiendo a categorías gramaticales distintas de las recogidas por dicha norma. La categoría gramatical en la que nos centraremos es la de los verbos. Se revisan, así, las posibles ventajas que pueda tener la inclusión de formas verbales como términos documentalmente relevantes a la hora de contextualizar la información y automatizar la construcción de los tesauros.

Van Slype [1] amplía el número de relaciones presentes de facto en los tesauros hasta cinco, incluyendo las relaciones que él denomina de pertenencia y de equivalencia interlingüística. Ahora bien, en la propuesta presentada por Van Slype aparece un sexto grupo de relaciones que él denomina no-clásicas por no estar presentes en la mayor parte de los tesauros en uso [2]. Este grupo de relaciones englobadas por él bajo el epígrafe único no-clásicas son: parte-todo, colocación, relación paradigmática, taxonomía y sinonimia.

Si analizamos con detenimiento estas últimas relaciones, podemos concluir que no son más que reformulaciones de relaciones que ya estaban contempladas en las relaciones que Van Slype denomina clásicas. Así, por ejemplo, la relación no-clásica parte-todo, que él ejemplifica con "cuerno PART vaca", no deja de ser un caso concreto de la relación jerárquica partitiva, en la que un descriptor es una parte particular de otro descriptor superior. En los ejemplos de este autor, "Francia y Europa" o "motor y vehículo", responden ambos a la pregunta ¿es uno parte de otro? [3].

Las relaciones que están experimentando una mayor expansión son las de asociación. La expansión se está produciendo a dos niveles: por una parte se están definiendo nuevas relaciones de asociación y, por otra, se están considerando como relaciones de asociación algunas relaciones que antes eran consideradas sólo de jerarquía.

En la actualidad, esta duplicidad esbozada por Van Slype está siendo concretada para tesauros automatizados. El trabajo de D. Tudhope, H. Alani y C. Jones [4] sobre la expansión de las relaciones en los tesauros automatizados actuales es una buena muestra de ello. Para estos autores, las relaciones que están experimentando una mayor expansión son las de asociación. La expansión se está produciendo a dos niveles: por una parte se están definiendo nuevas relaciones de asociación y, por otra, se están considerando como relaciones de asociación algunas relaciones que antes eran consideradas sólo de jerarquía. Basándose en un estudio presentado por la American Library Association (ALA) [5], Tudhope, Alani y Jones hablan de nueve grandes

2. *Documentación – directrices para el establecimiento y desarrollo de Tesauros monolingües (parte I)*

3. *La norma ISO 2788 (1986), en su apartado 8.4.1., resume a dos los posibles tipos de términos que puede conectar la relación asociativa:*

   a) Aquellos que pertenecen a la misma categoría;

   b) Aquellos que pertenecen a categorías diferentes.

   Dentro de los que pertenecen a la misma categoría, señala como único grupo el de aquellos términos emparentados a través de su significado (apartado 8.4.2.1.). En cuanto a los que pertenecen a categorías diferentes, la enumeración de los tipos se extiende a diez: una disciplina o campo de estudio y los objetos o fenómenos estudiados, una operación o proceso y su agente o instrumento, una acción y el resultado o producto de tal acción, una acción y su sujeto pasivo, conceptos y sus propiedades, conceptos relacionados con sus orígenes, conceptos ligados por una dependencia causal, objetos y sus contraagentes, conceptos y unidades de medida, frases sincategoremáticas y los sustantivos implicados.

grupos de relaciones de asociación que engloban las once que presenta la norma ISO 2788 (1986),[3] y que, con sus correspondientes subtipos, aumenta el número de forma considerable el número de posibles relaciones de asociación de un tesauro.

Estos diez subtipos de relaciones de primer nivel son:

1) Ideas combinadas.

2) Términos relacionados conceptualmente.

3) Contigüidad.

4) Relaciones asociativas por definición.

5) Relaciones asociativas con diferente jerarquía o facetadas.

6) Relaciones asociativas traslapadas por significado.

7) Relaciones asociativas con idéntica jerarquía.

8) Cuestiones de finalidad.

9) Relaciones asociativas sin especificar.

Los niveles llegan hasta seis en profundidad, lo que da un número total de relaciones de asociación de cientoveintitrés. Esto supone un aumento de cientodoce relaciones de asociación. Por otra parte, un estudio de las relaciones de jerarquía que ofrecen estos dos autores, basado también en los estudios de la ALA, nos lleva a la conclusión de que muchas relaciones que eran recogidas en el conjunto de las denominadas de jerarquía han pasado a ser clasificadas como relaciones de asociación. Algunas de ellas coinciden con las que Van Slype llamaba no-clásicas, como por ejemplo las relaciones del tipo parte-todo, que están empezando a ser consideradas como de asociación, aunque siguen apareciendo como de jerarquía, y son las que se duplican en la nueva caracterización de la ALA (aparecen como relaciones de asociación y de jerarquía, como por ejemplo las de parte-todo de sistemas u órganos anatómicos).

El principal problema que enfrentan los tesauros clásicos en la actualidad no es otro que el de la obsolescencia.

Este cambio en las relaciones de un tesauro viene dado, a nuestro entender, por el principal problema con el que se enfrentan los tesauros clásicos en la actualidad que no es otro que el de la obsolescencia. Y la razón fundamental de la obsolescencia de los tesauros hay que buscarla, a nuestro juicio, en la nota característica de los tesauros tradicionales: la estaticidad de la indización y de las relaciones establecidas que les es propia[4] Las estructuras que se encuentran definidas en un tesauro clásico hacen que sólo pueda recuperar los documentos que posean aquellos términos y relaciones estáticas con las que trabaja y que han sido definidas a priori. Esta limitación, que conlleva la obsolescencia, provoca el silencio o el ruido en los casos en que la búsqueda se haga sobre los mismos términos y relaciones predefinidas, pero en los que hayan cambiado su significado debido a nuevos usos. Esta clase de problemas surgían ya cuando el universo de discurso documental es el que vamos a calificar de clásico,[5] esto es, cuando se resume a unos tipos de documentos concretos muy bien encuadrados en un ámbito cultural y científico, o en ambos, perfectamente delimitado. En este caso, el de los documentos clásicos, hay que señalar, además, que, fuera cual fuera la temática concreta de los documentos tratados y su forma de presentación, siempre contaban con una presentación física (tridimensional), que permitía su localización espacio-temporal en unas coordenadas concretas. Es decir, los tesauros clásicos se diseñaban pensando en un universo documental muy bien parcelado temáticamente (salvando excepciones) y que contaba con una plasmación física concreta de los documentos a los que se orientaba.

4. Hemos decidido dejar fuera de este contexto otro tipo de problemas que pueden influir en la eficacia de un tesauro por referirse más al ámbito de los recursos humanos como pueden ser: la especialización, o falta de ella, de aquellos que llevan a cabo un tesauro concreto, los errores tipográficos y conceptuales debidos al factor humano, por mencionar algunos ejemplos.

5. Llamamos documentos clásicos a los documentos existentes antes de lo que se empieza a conocer como la era de la Sociedad de la Información, que va acompañada por el fenómeno de la globalización.

6. Documentos como pueden ser páginas web, que sólo tienen vida mientras que están en ella, pero que desaparecen definitivamente si su autor o autores deciden en un momento dado retirarlas de la web.

El universo documental actual ha cambiado sensiblemente con la aparición de la denominada Sociedad de la Información

Sin embargo, el universo documental actual ha cambiado sensiblemente con la aparición de la denominada Sociedad de la Información, y lo ha hecho en los dos aspectos señalados. Por una parte, la creciente especialización en los diferentes ámbitos culturales, en concreto los que están relacionados con los campos científicos surgidos desde la misma rama común, hace mucho más difícil la tarea de distinguir en clases diferentes los textos que son de un área científica y los de otras que se encuentran muy próximas a ella [6]. Además, los contenidos de determinadas áreas como las relacionadas con la Informática, están ofreciendo documentos expresados en lenguajes mucho más abstractos que los usados hasta el segundo tercio del siglo XX, con gran cantidad de gráficos y formatos de diseño, en resumen, menos alfabetizados. Por otra parte, las formas de presentación de los documentos (y no nos referimos sólo a los diferentes soportes posibles) ya no se limitan a las coordenadas espacio-temporales, sino que existen documentos (muchos de los contenidos en la red Internet) que no son localizables mediante estas coordenadas.[6] La consecuencia directa de estos cambios es la imposibilidad de llevar a cabo un tesauro específico, fiable y actualizado, que cumpla bien las funciones de indización precisa, pero sobre todo, que sea útil en el momento de la recuperación. Así pues, los problemas concretos que presenta un tesauro clásico se traducen en la práctica en dos graves inconvenientes:

1) *Dificultad para describir campos con un nivel de abstracción alto* (el ámbito de la Informática, por ejemplo), o en aquellos en los que, por el contrario, los documentos no estuvieran estructurados en ninguna medida (como puede ser el caso de los textos de libre formato).

2) *Un elevado costo, ya sea en la creación o en el mantenimiento*, que sólo se podría eliminar con la automatización o semiautomatización de ambas tareas. Lamentablemente, esta automatización difícilmente puede ser llevada a cabo de forma eficaz y eficiente desde la caracterización actual de los tesauros estáticos, debido a la riqueza de las estructuras semánticas de los textos (habría que definir una infinidad de nuevas relaciones cada vez que nos enfrentásemos a nuevos textos).

El dinamismo en la producción de información, propio de esta nueva Sociedad, cuenta con una nítida proyección epistemológica y apunta directamente al conocimiento

Hay que tener en cuenta, además, que el dinamismo en la producción de información, propio de esta nueva Sociedad, cuenta con una nítida proyección epistemológica y apunta directamente al conocimiento. Si bien durante la modernidad el acceso a la observación y la comprensión del mundo eran objetivas, es decir, se encontraban fuera del individuo, en el postmodernismo, sin embargo, tanto una como otra no existen fuera del individuo, ya que la tecnología permite al individuo diseñar su propio acceso a la información y, en última instancia, al conocimiento. La repercusión es inmediata para la organización del conocimiento, pues desaparecen las referencias absolutas y, por consiguiente, la clasificación y ordenación de temas deben referirse más a conveniencias, o posibles conveniencias, constituyendo de este modo una realidad que satisface en un momento y para una finalidad determinados.[7]

Se produce, lógicamente, una disarmonía entre el dinamismo en la creación de la información -y, por ende, del conocimiento- y los actuales lenguajes de clasificación del conocimiento, fundamentados en campos académicos normalizados. La clasificación por disciplinas es un sistema de clasificación del

7. Arguye el dinamismo del conocimiento en el postmodernismo Miska, F. L. The DDC, The universe of knowledge and the postmodern library. Forest Press Albany (NY), EE. UU., 1998.

conocimiento íntimamente referido a un método de investigación racionalista y pragmático. Este sistema está perdiendo utilidad debido, en gran parte, a los nuevos métodos y formas de producir, diseminar y usar la información que constituye, en última instancia, el receptáculo natural del conocimiento. Las nuevas comunidades discursivas cooperan en la producción de documentos que no se adaptan a las estructuras académicas [7], en tanto que las búsquedas se orientan cada vez más a un material no disciplinar. La automatización y la técnica hipertextual han dado respuestas en la recuperación documental, pero su eficacia está determinada por la necesidad de una nueva clasificación del conocimiento para optimizar la información. A este respecto, se ha desarrollado el concepto de disciplinaridad compartida en la creación del conocimiento, lo cual trae consigo que, para autores como D. W. Langridge [8], la clasificación deba distinguir entre disciplinas (formas de conocimiento) y fenómenos (objetos de conocimiento).

Las formas verbales como conceptos dinámicos

Como se comentó al inicio del presente documento una de las corrientes principales a la hora de denominar nuevas relaciones en tesauros se basa en el empleo de formas verbales en los tesauros. Aunque la carga conceptual de los textos recaiga en los sustantivos, éstas constituyen los conectores naturales de esos conceptos de los que los sustantivos son depositarios, y son las que dan un sentido concreto a esa carga conceptual,[8] que estaría incompleta si las formas verbales no completasen su contenido. Es decir, aunque el contexto lo aporten los sustantivos de forma habitual, las formas verbales son las que declaran las acciones y, consecuentemente, las que dan una orientación concreta a un determinado texto, cuando no son las que dotan al texto de la carga conceptual necesaria para clasificarlo.

Una de las corrientes principales a la hora de denominar nuevas relaciones en tesauros se basa en el empleo de formas verbales

La inclusión de los tesauros de formas verbales que complementen, sin eliminarlos, a los tesauros estáticos tradicionales parece la solución que mejor recoge la problemática actual. Como descriptores, las formas verbales presentan evidentes ventajas en determinados casos como, por ejemplo, a la hora de la realización de búsquedas de documentos en los que se recojan acciones y para la indización de materiales especiales. Ejemplos concretos de estos usos son: la indización de imágenes de vídeo mediante la identificación del verbo en gerundio que expresa la acción, o la utilización de formas verbales en los diagramas de clases tanto en la denominación de los métodos como en las asociaciones de clases. Otro ejemplo clarificador puede verse con los documentos de análisis de requisitos, este tipo de documentos cuentan con estructuras lingüísticas sencillas unidas por formas verbales (por ejemplo, afirmaciones como "El parámetro B debe ser mayor que 18"), y en las que lo importante es cuál es la conexión que se establece entre los elementos relacionados, y no ellos mismos.

La unión de un tesauro estático clásico de sustantivos con uno dinámico de formas verbales permitiría una indización automática más flexible

8. *Desde el punto de vista lingüístico toda proposición debe tener un verbo, es más, todo verbo, incluso aunque esté aislado, constituye una proposición, y no lo es la expresión que carezca de él.*

Por otra parte, la identificación del papel de una asociación mediante un verbo permite un abanico de relaciones mucho más adaptable a dominios concretos. Además, existe la posibilidad, en caso de realizar una indización automatizada, de grabar para su posterior uso ciertos modos y tiempos verbales como matices de la relación con la entidad, que pueden ser muy útiles en las tareas de identificación y localización de un documento a través de la lematización. Este aserto halla su más concreta constatación en la documentación científica, donde se emplean determinadas formas verbales para plantear la hipótesis (condicional), el estado de la cuestión (presente), método (futuro, potencial), conclusiones (pasado), como también un jugoso empleo del gerundio en la indización para los documentalistas de unidades de información audiovisual, tal como ya se señaló.

En resumen, esta nueva fórmula, que supone la unión de un tesauro estático clásico de sustantivos con uno dinámico de formas verbales, permitiría, a nuestro entender, una indización automática más flexible, es decir, la dinamización del tesauro eliminaría en gran medida los problemas en la representación de algunos dominios y aumentaría, al mismo tiempo, la precisión y eficacia de los mismos.

Justificación documental de la utilización de formas verbales en la recuperación automática de información

Dado que las principales funciones de un tesauro son la indización y la recuperación desde el punto de vista de la carga conceptual de los documentos, parece lógico que la atención a la hora de su diseño debe centrarse en cómo conseguir que el tesauro detecte la carga conceptual relevante en cada documento, atendiendo a su estructura lingüística en sus cuatro vertientes: léxica, morfológica, sintáctica y semántica.

Desde una perspectiva puramente documental, se podrían dividir en dos tipos de redes los resultados de la clasificación automática de la información. Por un lado, estarían las redes que crean las relaciones con carácter permanente, esto es, lo que se representa en los tesauros clásicos. En este tipo de red tiene mucha importancia contar con un número limitado de relaciones de asociación, equivalencia y jerarquía. Estas relaciones son percibidas por todos como obvias en determinado dominio (o, si se prefiere, son inherentes al conocimiento compartido por una comunidad determinada) y son analizados, independientemente del vocabulario empleado en los corpora del dominio (ya que son creados a priori con respecto al proceso de indización). Estas relaciones que denominamos permanentes son las que aparecen en los tesauros clásicos de manera habitual.

Por otra parte se encuentra la red de relaciones circunstanciales, en la que sí es importante tener un número elevado de formas verbales para representar todas y cada una de las relaciones que se puedan dar. Las relaciones que tenemos en una red de relaciones circunstanciales son, principalmente, del tipo "asociado con", y, en este caso, necesitamos, además, un conjunto de subtipos dependientes del dominio para subdividirlas, esto es, necesitamos una serie de especificaciones dadas por el dominio concreto y que sólo rigen en él. Es decir, el por qué de la necesidad de un gran conjunto de tipos de relaciones circunstanciales (representados por formas verbales) se debe a que el contenido (la semántica) de cierto documento distinto al de los recogidos hasta ese momento es, por su misma novedad, impredecible. Una red de relaciones circunstanciales es la herramienta idónea para representar el contenido de una colección documental (y para, posteriormente, facilitar la recuperación).

Actualmente los trabajos se orientan a utilizar las gramáticas semánticas y sintácticas para establecer relaciones

De cara a la automatización, es la red circunstancial (la correspondiente a los subtipos del dominio) la determinante de la mayoría de sistemas automáticos de clasificación. En estos sistemas se selecciona un corpus de documento y sobre la base de la frecuencia de coaparición de las palabras en los documentos, y entre los documentos, se hace una clasificación de tipo estadístico (redes neuronales, kmeans axiales, coocurrencia de términos, algoritmos genéticos, etc). Es decir, se parte de la indización de un corpus documental del dominio para construir el tesauro. De hecho, tradicionalmente se han creado redes de relaciones no sólo circunstanciales mediante herramientas estadísticas; actualmente los trabajos en ese sentido se orientan a utilizar las gramáticas semánticas y sintácticas para establecer relaciones, sin que esto suponga que no se pueda mejorar los resultados o acotar los listados de términos con herramientas estadísticas clásicas.

Una cuestión que queda pendiente es si es posible pasar de un tesauro de relaciones permanentes a uno de relaciones circunstanciales, basándonos exclusivamente en parámetros estadísticos. Parece ser que sí, pero siempre que se parta de un conjunto de documentos lo suficientemente grande, pero esto es, precisamente, lo que se pretende eliminar con la automatización.

Antecedentes del uso de formas verbales en los tesauros

Desde un punto de vista histórico, nuestra propuesta no constituye el primer intento en el que se usan las formas verbales como elemento en la recuperación automática o manual de la información; así como tampoco es la primera vez que se usan para la fundamentación teórica del tipo de herramientas que deben ser usadas en estas tareas. Ya en la década del 60 y del 70 existen algunos antecedentes en la aplicación automática y manual de los verbos para la mejora de la recuperación de información y como elementos claves a tener en cuenta en la fundamentación teórica.

Entre los antecedentes estrictamente teóricos que proponen al verbo como elemento principal en la indización y la recuperación de los documentos cabe destacar la propuesta de Alan L. Tharp. En un escrito de 1973 titulado Using verbs to automatically determine text descriptors, Tharp [9] hacía una defensa, desde el punto de vista teórico, de este uso de los verbos en el contexto correspondiente a las tareas de recuperación automatizada de la información. Según Tharp, los métodos de indización automática que se basan exclusivamente en estadísticas, permutaciones, citaciones o asociaciones no permiten la exactitud en la búsqueda y la recuperación, puesto que lo hacen atendiendo a criterios extrínsecos. A su juicio, este tipo de métodos indizan sin tener en cuenta el contexto de cada uno de los términos.

Una cuestión que queda pendiente es si es posible pasar de un tesauro de relaciones permanentes a uno de relaciones circunstanciales, basándonos exclusivamente en parámetros estadísticos

La idea de fondo que mantiene Tharp a lo largo de su argumentación es que un buen sistema de recuperación de la información debe seguir los esquemas de memoria del ser humano. Y, según sus propias palabras, "la simulación de la memoria humana conduce a usar el verbo como el elemento significativo en la determinación de los descriptores del texto" [10, p. 247]. Esta afirmación se apoya, a su vez, en estudios psicológicos que demuestran que es una característica de la memoria humana el ser imprecisa para recuperar de manera estática, y, sin embargo, muy fiable en la recuperación del movimiento, esto es, de acciones.

Tal como sostiene Tharp, conviene recordar que todos los sistemas de indización, clasificación y recuperación automática se basan en la medida de diversos factores (estadísticas, permutaciones, citaciones y asociaciones, por ejemplo). A juicio de Tharp, el hecho de que los verbos significativos, es decir, los que transmiten una mayor cantidad de información contextual, aparezcan en un número muy pequeño de veces, los haría desaparecer de la lista de posibles términos del documento si se utilizan los anteriores parámetros, lo cual resulta contraproducente para recoger la carga conceptual

9. *Un completo estudio se encuentra en Halliday and Hasan, M. A. K. Language, context, and text: aspects of language in a social-semiotic perspective. Universidad de Oxford, 1989.*

completa del mismo. Siguiendo los mismos estudios propuestos para los sustantivos en el caso de los verbos (y sorteando los problemas de las formas flexionadas de los verbos), tendríamos que esos verbos con poca frecuencia de aparición, pero muy representativos desde el punto de vista contextual, dejan de ser candidatos para representar al texto como los identificadores del mismo a la hora de la búsqueda y recuperación del documento. Es lógico pensar que cualquier tipo de filtro o normalización implica una pérdida de semántica, por ejemplo, podemos normalizar los términos niñas, niña, niño y niños a una misma forma canónica, por ejemplo, niño. Pero ¿qué ocurre si en el futuro queremos sólo recuperar documento que traten de niñas pero no de niños? Es por el mismo hecho por lo que hacer desaparecer en la indización términos con carga conceptual relevante, en este caso los verbos, puede limitar en el futuro las posibilidades de búsqueda y recuperación.

Un ejemplo de plasmación concreta de este tipo de desarrollos automatizados o semiautomatizados se puede encontrar en el SYNTOL [11]. El SYNTOL fue un estudio auspiciado por la Comunidad Europea de la Energía Atómica (EURATOM) para la elaboración de un tesauro especializado, que intentaba sortear la caducidad de los lenguajes de indización del tipo de los tesauros al uso en ese momento [12]. La implantación del SYNTOL fue un fracaso en su momento por su carestía; ahora bien, el elevado coste estaba relacionado directamente con la limitación de los ordenadores con los que se contaba en ese momento. Los equipos actuales abaratarían notablemente el gasto en este tipo de implementaciones, y esto tanto desde el punto de vista crematístico como por el tiempo que las máquinas necesitarían emplear actualmente para llevar a cabo las rutinas de indización y recuperación.

Sobre la utilización de formas verbales aplicada a la contextualización de la información

La utilización de formas verbales aplicada a la contextualización de la información ha sido estudiada en tres vertientes concretas: determinación de géneros, determinación de la tipología de las secciones del documento y la determinación temática.

En cuanto a la determinación de géneros, Swales [13] caracterizó los géneros según los formas verbales presentes en cada documento. Se centró en dos factores: cuáles son los verbos prototípicos y en qué flexión (tiempo/modo/número) se encuentran más frecuentemente. Por ejemplo, en un estilo no académico destaca la disminución de formas verbales en pasiva y presente simple aumentando la del pasado simple.[9] Si es posible esta caracterización es de suponer que el proceso contrario también es posible. Es decir, buscando ciertos verbos y flexiones verbales podemos contextualizar el género en el que se produce, lo que solucionaría en gran medida la problemática de los documentos que se pueden localizar en Internet, que no es otro que la mezcla de distintos géneros con calidad variable.

En el caso del vocabulario científico, por ejemplo, prevalecen los verbos de tipo informativo como: demostrar, establecer, argumentar,... Así como también son frecuentes en dicho ámbito los verbos negativos que nos indican las carencias que intentamos salvar con la documentación: faltar, subestimar, adolecer, eludir, etc.

En lo relacionado con la determinación de la tipología de las secciones del documento, con un enfoque similar se ha determinado el posicionamiento en las distintas secciones de los documentos. Un buen ejemplo es el Heslot [14], quien diseñó el esquema de distribución de formas verbales que aparece en la tabla 1.

Tanto la determinación de géneros como de secciones de un documento difieren en su modo y tenor dependiendo del campo [15] al que pertenezca el documento, lo que nos lleva, nuevamente, a la necesidad de aplicar también una red semántica en la indización y establecimiento de relaciones de cara a la recuperación. En este punto conviene recordar que ya se han empleado este tipo de herramientas en la contextualización de la información, como puede ser el intento de Haas [16] para determinar de modo automático artículos de carácter experimental, sin embargo, tampoco el trabajo de Haas cuenta con un tesauro de formas verbales que complemente a los clásicos de sustantivos.
Por último, en la determinación temática una de las vías de investigación más exitosas en los últimos años es WordNet [17]. La forma de afrontar el problema es utilizar los términos raíces de los verbos para acotar tanto el género como el tema [18]. Esta red semántica de WordNet presenta, sin embargo, los mismos problemas que señalábamos en el caso anterior, es decir, no cuentan con un tesauro de formas verbales que sirva de complemento al de

Tabla 1. Esquema de distribución de formas verbales de J. Heslot

CARACTERÍSTICA	INTRODUCCIÓN	MÉTODOS	RESULTADOS	DISCUSIÓN	AUTORES
Comentarios autor	Alto	Muy bajo	Muy bajo	Alto	Adams Smith
Pasiva	Bajo	Alto	Variable	Variable	Heslot
Pasado	Bajo	Alto	Alto	Bajo	Heslot
Presente	Alto	Bajo	Bajo	Alto	Heslot
Información	Alto	bajo	Bajo	Alto	West

sustantivos, ya que se limita a ser un tesauro de verbos sin ninguna relación semántica con el de sustantivos.

Papel de la categoría de las formas verbales en los tesauros

Aunque en la ISO R1087 se comenta que "... los conceptos pueden ser no sólo cosas (expresadas mediante sustantivos), sino que en un sentido más amplio, también se puede tratar de cualidades (adjetivos y sustantivos), de acciones (como por ejemplo, verbos y sustantivos) e incluso lugares, relaciones o situaciones (expresadas mediante adverbios, preposiciones, conjunciones y sustantivos) [19, p. 129]", no es menos cierto que, según la ISO 2788 (1986), en la construcción de un tesauro se deben evitar los usos como términos de indización de determinadas categorías gramaticales. Entre las categorías excluidas se encuentra la de los verbos [20], así como de adjetivos aislados –salvo determinadas excepciones-[21], y adverbios [22].

En cuanto a lo que atañe a la categoría gramatical de los verbos es preceptivo, según la segunda de las normas citadas, que se "presenten como sustantivos o formas verbales sustantivadas [23, p. 10]". La interpretación estricta de la anterior afirmación ha sido la que más se ha extendido y, como consecuencia, el resultado es que las formas verbales sólo han sido consideradas como términos de cara a la indización y la recuperación cuando funcionan como sustantivos en el sentido gramatical. Sin embargo, siguiendo la norma ISO R1087, el tipo de conceptos que representan los verbos (formas verbales) es el de las acciones, que pueden ser concretadas en sustantivos o verbos, según dicha norma. Los sustantivos que responden a este tipo de conceptos, el de las acciones, suelen ser los que comparten su semántica con un verbo, de quien la reciben. Por ejemplo, el sustantivo "navegación" recibe su carga semántica del verbo "navegar", "aceleración" de "acelerar", y así sucesivamente. Así pues, parece que nos encontramos frente a tipos conceptuales distintos, a los cuales hay que tratar, a nuestro juicio, desde su diferencia, y no como se ha hecho tradicionalmente, esto es, asimilándolos todos a la categoría conceptual de los sustantivos.

Mantener la idiosincrasia de los conceptos que representan acciones conlleva la inclusión en el tesauro de la categoría que los refleja en los documentos, sean estos sustantivos o formas verbales

Mantener la idiosincrasia de los conceptos que representan acciones conlleva la inclusión en el tesauro de la categoría que los refleja en los documentos, sean estos sustantivos o formas verbales. Así, por ejemplo, si nos encontramos con la frase "La aceleración del cohete espacial ha sido correcta", tendremos que buscar la carga conceptual de la misma, además de en el sustantivo que representa la cosa de la que trata la frase, "cohete espacial", en el sustantivo que representa la acción que concreta el contexto en el que se encuentra inmersa, "aceleración". Y, en última instancia, tendremos que remitirnos al verbo del que recibe su carga conceptual, "acelerar", que en este caso no aparece en la frase, es decir, el término "aceleración" no se refiere a una cosa, sino que está representando una acción. A nuestro entender, sólo así podremos recoger la carga conceptual de la frase, que quedaría incompleta si nos limitásemos al término "cohete espacial". Esta posibilidad de precisar aún más el contenido conceptual de un texto delimita mucho

más la caracterización tipológica del documento, y hace que la indización sea más exhaustiva y, consecuentemente, la recuperación más exacta.

Futuros desarrollos

La creación de este tesauro de formas verbales, como complemento al tesauro clásico de sustantivos pasa por un proceso que no es meramente automático, sino que incluye otro proceso de naturaleza más lingüística en el que se puedan encajar las categorías verbales a modo de relaciones facetables. Este planteamiento nos separa de la mayoría de proyectos de este tipo, ya que casi toda la literatura trata de localizar un conjunto de términos (a los que denominan clusters), que resultan relevantes por su frecuencia de coaparición en determinado contexto.

Existen tres grandes corrientes en la que la inclusión de formas verbales en la indización y construcción de los tesauros:

1) *Utilización de clasificaciones verbales,* para mejorar la recuperación. La mayoría de los trabajos se basan en las aportaciones hechas por Levin [24]. Por ejemplo, destacan los trabajos de Green [25] en el que se crea una ampliación de la red semántica WordNet con los verbos de Levin.

2) *Utilización de WordNet* [26]. Se trata de una red semántica multidisciplinar en inglés. Su calidad y disponibilidad lo han convertido en la herramienta idónea en lingüística durante los últimos años. Su uso ha estado centrado desde un principio muy vinculado a la desambiguación conceptual. Dentro de esta tendencia una de las líneas que está dando mejores frutos es el empleo para desambiguar mediante verbos. Un ejemplo típico de desambiguación se puede encontrar en Moldovan[10] en donde se esquematiza de la siguiente forma:[11]

a) Se selecciona de la frase todas las parejas de sustantivo-verbo.

b) Se escoge el significado más probable del término (subproceso que Moldovan denomina Desambiguación Terminológica):

i. *Las palabras de la frase se agrupan en parejas.*

ii. *Se buscan en WordNet los distintos significados de cada término.*

iii. *Se forman todos los diferentes pares de conceptos posibles.*

iv. *Se busca cada par en Internet.* Luego se ordenan los resultados según los conceptos más frecuentes, esto es, según el número de veces que en Internet nos aparezcan juntos dos conceptos determinados.

c) *Teniendo en cuenta los conceptos más frecuentes (paso b), se seleccionan todos los sustantivos de los "glosarios" de cada verbo* y sus descendientes jerárquicos.

d) *Teniendo en cuenta los conceptos más frecuentes, se seleccionan todos los sustantivos de los "glosarios" de cada sustantivo* y sus descendientes jerárquicos.

e) *Se calcula mediante una fórmula* los conceptos comunes entre los sustantivos del punto c y d.

f) *Se ordenan todas las parejas de conceptos de sustantivo-verbo* según el resultado de la fórmula.

3) *Utilización de las formas verbales de los documentos de determinado dominio* [27]. En esta vía se pretende automatizar la construcción de tesauros basados en las concurrencias de descriptores del tesauro con formas verbales en determinado corpus documental. Esta última vía es la que estamos desarrollando, aunque empleando algunas de las tareas señaladas en las dos anteriores, y es por eso por lo que las incluimos como futuros desarrollos.[12]

10. Moldovan también ha propuesto otros algoritmos de desambiguación basados en WordNet, véase Sanda M. Harabagiu & Dan I. Moldovan. Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text. En Natural Language Processing & Knowledge Representation. Ed. Lucja M. Iwánska and Stuart C. Shapiro. Menlo Park (CA): AAAI Press, MIT Press, 2000. pp. 301-333.

11. El autor denomina al método desambiguación semántica. Using WordNet and lexical operators to improve Internet searches. Moldovan-DI; Mihalcea-R. IEEE-Internet-Computing. 4) (1):34-43, Jan.-Feb. 2000. Con este sistema Moldovan afirma que obtiene una considerable mejora frente a Yarowsky (1995) que desambiguaba el 94% de los sustantivos (no otros términos) y a Stetina (1998) y este trabajo, desambiguar el 80% de todas las partículas.

12. En el próximo Congreso INFO 2002, que se celebrará en Cuba, presentaremos los primeros resultados de nuestro trabajo.

Referencias

1) Van Slype, G. *Los lenguajes de indización. Construcción y utilización de los sistemas documentales.* Fundación Sánchez Ruipérez. Madrid, España, 1991. 198 p.

2) Van Slype, G. Op. Cit., p. 58.

3) Van Slype, G. Op. Cit., p. 53.

4) Tudhope, Douglas, Harith Alani y Christopher Jones. *"Aumenting Thesurus Relationships: Possibilities for Retrieval"* [en línea]. Journal of Digital Information (1), febrero 2001. <http://jodi.ecs.soton.ac.uk/Article/v01/i08/Tudhope>. [Consulta: 27 de abril del 2001].

5) Appendix B (part 2) *Taxonomy of Subject Relationships compiled by Dee Michel with the assistance of Pat Kuhr* [en línea]. June 1996 draft (hierarchical display), en <http://ala.org/alcts/organization/ccs/sac/appendxb.html>. [Consulta: 8 de febrero del 2002].

6) Niiniluoto, Illka. *The emergence of Scientific Specialities: six models, Poznan Studies in the Philosophy of the Sciences and the Humanities* (44):211-223, 1995.

7) Clegg, S. R. y G. Palmer. *The politics of management knowledge.* Sage London, Inglaterra, 1996.

8) D. W. Langridge. *Classification, its kinds, systems, elements and applications.* Bowker Sauer London, Inglaterra, 1992.

9) Tharp, Alan L. *Using verbs to automatically determine text descriptors.* Information Storie and Retrieval (9):243-248, 1973.

10) Tharp, A., loc. cit., p. 247.

11) Cros, R. C., J. C. Gardin, y F. Lévy,. *L'automatisation des recherches documentaires. Un modèle general «LE SYNTOL»,* Gauthiers-Villars Paris, Francia, 1968, 260 p.

12) Cros, R. C., J. C. Gardin, y F. Lévy, *L'automatisation des recherches documentaires. Un modèle general «LE SYNTOL»*, pp. 5-6.

13) J. M. Swales,. *Genre analysis: English in academic and research settings.* Cambridge University Press, Cambridge, Inglaterra, 1990.

14) Heslot, J. *Tense and other indexical markers in the typology of scientific texts in English.* Hedt: 83-103, 1982.

15) Lavid, Julia. *Towards a text type taxonomy: a functional framework for text analysis and generation. Procesamiento Lenguaje Natural* (16):29-43, 1995.

16) Haas, S. W., J. Sugarman y H. Tibbo. *A text filter for the automatic identification of empirical articles.* Journal of the American Society for Information Science, 47(2):167-169, 1996.

17) Fellbaum, C. Wordnet. *An Electronic Lexical Database (Language, Speech and Communication).* MIT Press Cambridge (Massachusetts), 1998.

18) Klavans, Judith y Min-Y Kan,. *Role of Verbs in Document Analysis.* En Proceedings of the Conference, COLING-ACL. Montreal (Canada): Université de Montreal, 1998. pp. 680-686.

19) ISO R1087, citado por Yukio Nakamura: *"A Language for Knowledge Representation"*, en la revista Advances in Knowledge Organization, (4):127-133, 1994.

20) Norma ISO 2788 - UNE 50-106-90. *Directrices para el establecimiento y desarrollo de tesauros monolingües,* 1986, p. 10.

21) Norma ISO 2788 - UNE 50-106-90. *Directrices para el establecimiento y desarrollo de tesauros monolingües,* 1986, pp. 9-10.

22) Norma ISO 2788 - UNE 50-106-90. *Directrices para el establecimiento y desarrollo de tesauros monolingües,* 1986, p. 10.

23) Norma ISO 2788 - UNE 50-106-90. *Directrices para el establecimiento y desarrollo de tesauros monolingües,* 1986, p. 10.

24) Levin, B. *English verb classes and Alternations: a preliminary investigation*, University of Chicago Press Chicago, Ill, 1993.

25) Green, Rebecca, Lisa Pearl, Bonnie J. Dorr and Philip Resnik. *Mapping Lexical entries in Verbs Database to WordNet Senses.* En Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001), Toulouse, France, July 9-11, 2001.

26) Fellbaum C., op. cit.

27) Díaz Rodríguez, S. I. *Esquemas de representación de información basados en relaciones : aplicación a la generación automática de representaciones de dominios,* Tesis doctoral, Director, Juan Lloréns Morillo. Leganés: Universidad Carlos III de Madrid, Departamento de Informática, 2001.

Recibido: 8 de febrero del 2002.
Aprobado: 10 de marzo del 2002.

José Antonio Moreiro González

Departamento de Biblioteconomía y Documentación
y Departamento de Informática
Universidad Carlos III de Madrid
C/ Madrid, 126 28903 Getafe (Madrid) España
Correo electrónico: <jamore@bib.uc3m.es>.