Procesamiento del Lenguaje Natural, num. 29 (2002), pp.283-284 recibido 11-07-2002; aceptado 15-07-2002

Desarrollo de un sistema de recuperación para entornos de información dinámica: creación de un tesauro de verbos para implementar el estándar ISO/IEC 13250:2000

José Antonio Moreiro González
Universidad Carlos III de Madrid
Facultad de Documentación, 28903 Getafe
jamore@bib.uc3m.es

Juan Lloréns Morillo
Universidad Carlos III de Madrid
Facultad de Informática, 28911
llorens@ie.inf.uc3m.es

1 *Ficha del Proyecto*

- Título del Proyecto: Desarrollo de un sistema de recuperación para entornos de información dinámica: Tesauros de verbos, implementación del estándar ISO/IEC 13250:2000
- Entidad Financiera: Proyecto financiado por la CICYT (Comisión Interministerial de Ciencia y Tecnología) del Plan General para el Conocimiento. TIC 2000-2003.
- Grupos participantes: Departamento de Biblioteconomía y Documentación y Departamento de Informática de la Universidad Carlos III de Madrid.
  Investigador responsable y coordinador: Dr. José Antonio Moreiro González.
  Departamento de Documentación. Universidad Carlos III de Madrid. Facultad de Humanidades, Comunicación y Documentación. c/ Madrid, 126 – 28903 Getafe (Madrid). Telf. 91 6249238. jamore@bib.uc3m.es
- Investigador responsable del Departamento de Informática: Dr. Juan Lloréns Morillo.
  Escuela Politécnica Superior. Avenida Universidad, 30 – 28911 Leganés (Madrid).

2 *Resumen*

Se propone un método que permita la construcción automática de *Topic Maps* (ISO/ICE 12350-2000). Con este objetivo se propone la creación de una clasificación de elementos del lenguaje natural que conduzcan a la identificación semiautomática de determinada *association type* del *Topic Maps*. En un primer momento, los tesauros (ISO 2788:1986) serán considerados a efectos de este trabajo como una simplificación de los *Topic Maps*. La construcción automática de los tesauros será la etapa anterior a la construcción de *Topic Maps* y servirá para valorar la clasificación de los elementos del lenguaje que aportan semántica.

3 *Objetivos*

Uno de los objetivos del proyecto se resume en la creación automática de *Topic Maps*. Una de las características más destacables de este estándar es la utilización de *association types*. Las *association types* son tipos de relaciones definidas por la categoría verbal junto con otras partículas anexas a ella. Para este fin, otro de los objetivos propuestos es una clasificación de *association types* y de las estructuras verbales que las definen, y por lo tanto, de relacionar conceptos unidos por otro concepto de carácter verbal, con la finalidad de aportar semántica para una representación del conocimiento de un dominio.

4 *Metodología*

El primer estadio de la metodología se centra en el estudio y clasificación semántica de las formas y estructuras verbales que representan las *association types*. El procedimiento será manual, tomando como base del estudio un corpus de un dominio específico. En este apartado se tendrán en cuenta diferentes propuestas recientes referentes a relaciones entre términos. Y también serán valorados los estudios previos sobre las clasificaciones de verbos

El segundo estadio está dedicado a la extracción del vocabulario contenido en los documentos de partida. Para este fin se ha creado una herramienta automática capaz de identificar el vocabulario, extraer, filtrar y referenciar la terminología representativa del domi-

ISSN 1135-5948

nio tratado. Una vez seleccionado los términos y las estructuras verbales (que los relacionan), la herramienta procede a establecer las *association types* correspondientes. Es decir, mediante las estructuras verbales se establecen tipos de relaciones entre los términos de los argumentos, creando la taxonomía y representación conceptual del dominio, similar a una estructura de Tesauro.

La herramienta se encarga de analizar las relaciones extraídas mediante axiomas que deben cumplir los elementos de la ontología (en la web semántica). Paralelamente, se contempla utilizar un algoritmo similar a tf-idf para seleccionar aquellas asociaciones más prometedoras.

La finalidad de esta herramienta es que mediante esta clasificación sea posible obtener automáticamente tesauros, y *Topic Maps*, a través del análisis de documentos de determinado dominio.

Dado el estado del proyecto su realización permite concluir que las metodologías encaminadas a facilitar la organización automática del conocimiento y la clasificación de los recursos asociados va a tener un impacto creciente en los próximos años.

Dentro del ámbito concreto del proyecto, tiene especial interés:

- El mayor número y riqueza semántica de las relaciones permitirá mejorar la recuperación al reducir la ambigüedad de los términos de las consultas.

Las herramientas creadas facilitarán la construcción automática de *Topic Maps* y de Tesauros. Este punto tendrá especial relevancia en el entorno de Internet haciendo posible la construcción de la web semántica.

5 Resultados

El proyecto comenzó en febrero de 2001 y finalizará en junio de 2003. La situación actual del proyecto se encuentra en el estudio de la clasificación de relaciones y las estructuras verbales que las establecen.

Se ha desarrollado una herramienta que indiza con los correspondientes módulos de normalización, generación de términos compuestos y filtrado terminológico. No obstante se siguen realizando estudios para perfeccionar cada uno de los módulos. Así como, se encuentran en continua investigación los complejos procesos de identificación de relaciones y de filtrado de las mismas.

6 Trabajos Futuros

En las futuras fases del proyecto se perfeccionarán las herramientas de extracción de terminología y de generación y filtrado de relaciones y se determinará la clasificación definitiva de las relaciones y las formas y estructuras verbales que las identifican.

Para una optima visión de los resultados obtenidos se compararán los resultados de recuperación de tesauros creados manualmente con los *Topic Maps* creados a partir de los documentos previamente indizados. El método para evaluar las clasificaciones esta previsto realizarlo mediante colecciones de evaluación (*test collections).*

Bibliografía

ISO/IEC 13250. 2000. *Information technology – SGML Applications – Topic Maps*, ISO, Geneva.

ISO-2788. 1986. *Guidelines for the Establishment and Development of Mono-lingual Thesauri.* International Organization for Standardization, Second edition -11-15 UDC 025.48. ISO, Geneva.

ISO-5964. 1985. *Guidelines for the establishment and development of multilingual thesauri.* International Organization for Standardization, ISO, Geneva.