MESIA: Recuperación de documentos en el sitio Web de la Comunidad Autónoma de Madrid utilizando recursos lingüísticos

Paloma Martínez[1], Paloma González[2], Pablo Sánchez[2] y Ana García-Serrano[2]

[1] Grupo de Bases de Datos Avanzadas
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es

[2] Grupo ISYS-Hermeneumática
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
{agarcia,pglez,psanchez}@isys.dia.fi.upm.es

Resumen

Se presenta el prototipo denominado MESIA[1] (Modelo Computacional para Extracción Selectiva de Información de textos cortos) cuyo objetivo es mejorar los resultados de la búsqueda de documentos en la Web de los buscadores tradicionales basados fundamentalmente en métodos estadísticos. Para ello se amplía la consulta con diferentes rasgos semánticos y estructurales obtenidos a partir del análisis del contenido de las páginas por el sistema que integra recursos disponibles para el tratamiento automático del castellano (Aries y EWN-español) .

El dominio de aplicación es la recuperación de información en el sitio Web de la Comunidad Autónoma de Madrid (CAM).

Funcionalidades

El sistema permite *transformar* la consulta del usuario en LN en una consulta formal que el buscador convencional de la CAM pueda ejecutar; *extender* los términos significativos de la consulta formal utilizando conocimiento lingüístico. Esta ampliación se hace de dos formas: mediante variaciones morfológicas y términos sinónimos o semánticamente relacionados.

Existe un módulo wrapper que se encarga de analizar las páginas HTML devueltas por el Buscador de la CAM con el fin de extraer la información textual que contienen y que será tratada posteriormente por el módulo analizador de cadenas significativas. Su funcionamiento se basa en un parser que utiliza una gramática de las páginas que describe las diferentes secciones y subsecciones interesantes de una página.

En el análisis lingüístico parcial de las cadenas significativas del texto se utilizan un conjunto de Patrones Semánticos que guían el análisis basado en expectativas según la terminología y palabras clave del dominio. Estos patrones hacen uso también de un *shallow parser* que lleva a cabo una segmentación superficial de determinados segmentos de texto Martínez y García-Serrano (1998). Este análisis produce una estructura de rasgos semánticos que describen superficialmente el texto de una página. Las estructuras de rasgos obtenidas para las páginas resultado de una búsqueda se envían tanto a un gestor de conocimiento extraído para almacenarlas para futuras consultas como a un presentador de resultados para organizarlas y visualizarlas al usuario.

La consulta introducida por el usuario se almacena junto con la consulta generada por MESIA en la base de datos de documentos que contiene, además, información estructural de las páginas devueltas por el buscador (título, párrafo, links, etc.) junto con su formato XML, orden de relevancia de las páginas según diversos criterios, etc.

Por último, el Gestor del Conocimiento extraído se encarga de gestionar la Ontología del Dominio en la que se van encajando las estructuras de rasgos obtenidas tras el análisis de los documentos; esta Ontología es una especificación formal y consensuada del vocabulario utilizado para describir los conceptos de un determinado dominio y contiene las direcciones URL enlazadas por un conjunto conceptos del dominio junto con sus rasgos semánticos.

Además, MESIA también puede recibir directamente una consulta formal para buscar la respuesta en la BD de Documentos Clasificados

[1] Este trabajo está financiado por el Proyecto CAM-07T/0017/1998

junto con la Ontología del Dominio sin necesidad de hacer la búsqueda en Web.

Por último, lleva un control de las consultas que los usuarios han realizado a lo largo del tiempo mediante la base de datos de Consultas Frecuentes.

Hemos comenzando los trabajos para evaluación del sistema. En los sistemas de recuperación de información se utilizan otras métricas además del tiempo y el espacio utilizado, que requieren una evaluación sobre la precisión del conjunto de respuestas obtenidas y que una vez realizadas nos permitirán aportar conclusiones sobre las ventajas de este tipo de sistemas para mejorar los resultados de las búsquedas en la Web.

Ficha Técnica

Hasta el momento se dispone de una primera versión del núcleo del sistema MESIA, que hace uso de recursos lingüísticos como es el caso de ARIES, un léxico y etiquetador morfológico para español, Goñi et al. (1997) y, acabamos de recibir el paquete EuroWordnet para el tratamiento de sinónimos en español, Gonzalo et al. (1998), con lo que su incorporación será inmediata (sustituyendo la base de sinónimos ad-hoc utilizada hasta el momento); para el diseño del sistema basado en el conocimiento en este primer prototipo se ha utilizado Java para el interfaz y acceso a las bases de datos (Access) y el entorno de programación lógica CIAO-Prolog, Bueno et al. (1999), que ha facilitado la integración del léxico y del módulo para el análisis basado en preferencias Martínez y García-Serrano(1998) así como el desarrollo de los diferentes módulos del prototipo.

Referencias

Bueno et al. (1999), F. Bueno, D. Cabeza, M. Carro, M. Hermenegildo, P. López, and G. Puebla (1999) ``The Ciao Prolog System: A Next Generation Logic Programming Environment, REFERENCE MANUAL'' The Ciao System Documentation Series Technical Report CLIP 3/97.1, The CLIP Group School of Computer Science Technical University of Madrid.

Gonzalo et al. (1998), J. Gonzalo, M.F. Verdejo, I. Chugur, Fernando López, Anselmo Peñas. "Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet". *Revista SEPLN,* nº 23, 1998.

Goñi et al. (1997), Goñi, J. M., González, J. C. y Moreno, A. ARIES: A lexical platform for engineering Spanish processing tools. *Natural Language Engineering, 3 (4)*, pp. 317-345.

Martínez y García-Serrano (1998), Martínez, P. and García-Serrano, A. A Knowledge-based Methodology applied to Linguistic Engineering. In R. Nigel Horspool Ed., *Systems Implementation 2000: Languages, Methods and Tools.* London: Chapman & Hall, pp. 166-179, 1998.