# Construcción de un Sistema de Recuperación Multilingüe en la Web

**Antonio Ferrández**
Universidad de Alicante
antonio@dlsi.ua.es

**Alfonso Ureña**
Universidad de Jaén
laurena@ujaen.es

**Víctor J. Díaz**
Universidad de Sevilla
vjdiaz@lsi.us.es

## *1 Ficha del Proyecto*

El proyecto ha sido financiado por el Ministerio de Ciencia y Tecnología en su programa de Fomento de la Investigación Técnica (PROFIT) con código FIT-150500-2002-416. La duración del proyecto es de 18 meses. Los grupos participantes son las Universidades de Alicante, Jaén y Sevilla, siendo el investigador responsable y coordinador el Dr. Antonio Ferrández Rodríguez, del departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante, con dirección Campus de San Vicente del Raspeig. Aptdo 99.03080 Alicante. España.

## *2 Introducción*

El proyecto se enmarca dentro del campo de la Recuperación de Información Multilingüe y la Búsqueda de Respuestas. Ambos campos surgen como una extensión de la Recuperación de Información clásica cuyo objetivo principal es obtener una relación de documentos en la que figura la información relevante solicitada por una petición o pregunta de un usuario. Mientras que la Recuperación de Información clásica trabaja sobre un único idioma, es decir, tanto la pregunta como los documentos sobre los que se busca la información se encuentran en un solo lenguaje, la recuperación multilingüe supone que tanto la pregunta como los documentos no necesitan estar en el mismo idioma. Los sistemas de búsqueda de respuestas plantean el problema bajo otra perspectiva en la que lo importante no es la flexibilidad en el idioma utilizado sino la calidad de la respuesta. De esta forma, la salida del sistema no debería consistir en una serie de documentos completos, sino el trozo de texto que se supone contiene la respuesta del usuario.

## *3 Objetivos*

El objetivo fundamental que se persigue en el proyecto es la construcción de un buscador de información multilingüe en el ámbito de los idiomas de la comunidad económica europea. Inicialmente, el sistema está centrado en dos idiomas, español e inglés, aunque se pretende ampliar el desarrollo para la incorporación de nuevos lenguajes en el futuro.

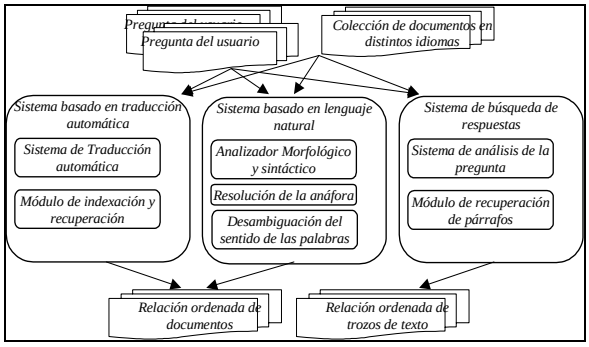


Figura 1 Arquitectura del sistema

El buscador que se está desarrollando pretende mejorar los sistemas tradicionales que trabajan sobre la web desde tres puntos de vista:

1. Se añadirá la capacidad de trabajar con varios idiomas. Es decir, con independencia del idioma que use el usuario para facilitar la pregunta, se devolverá una relación de documentos relevantes que también podrán estar en diferentes lenguas. Este proceso se realizará de forma transparente al usuario.
2. Se incluirán nuevas fuentes de información que los buscadores clásicos actualmente no contemplan, y que proceden del campo del procesamiento del lenguaje natural (PLN): análisis léxico, sintáctico, etc.
3. Se mejorará la calidad de la información a devolver, puesto que en lugar de producir como salida el documento completo, se

refinará dicha salida ofreciendo el trozo de texto donde se encuentra la información requerida por el usuario.

La figura 1 muestra la arquitectura global del sistema que está constituida por tres sistemas especializados en cada una de las tres mejoras descritas. Se ha perseguido en todo momento un diseño modular que permita la reutilización de los módulos entre los sistemas. Cabe destacar la presencia de un interfaz de usuario que recogerá la pregunta del usuario en cualquier idioma, la cual se enviará al sistema seleccionado junto con la colección de documentos en la que se realizará la búsqueda. Finalmente, se obtendrá el resultado que será procesado para presentarlo al usuario de la forma adecuada. Es decir, ya sea como una relación de documentos o como una relación de trozos de texto. Cada motor de búsqueda realizará el tratamiento adecuado de la pregunta.

Para ampliar datos sobre las características del proyecto es posible consultar la página web: http://www.dlsi.ua.es/projectes/srim/informacion.html donde se puede encontrar la descripción del personal que está trabajando en el mismo, junto con sus direcciones de contacto, una descripción del proyecto (tanto en inglés como en español), y las herramientas que se están creando durante el desarrollo del mismo.

## *4 Resultados*

El proyecto comenzó en febrero de 2002 y finalizará en septiembre de 2003. Hasta la fecha se ha procedido a la evaluación de herramientas que serán integradas dentro de los componentes del sistema, se han seleccionado los recursos lingüísticos que apoyarán la construcción de los componentes del sistema y se está procediendo al desarrollo de las restantes herramientas que conformarán el sistema. En las futuras fases del proyecto se pretende pasar a la integración y evaluación global del sistema. A continuación enumeramos los recursos y herramientas utilizadas y las herramientas desarrolladas:

1. Se ha diseñado la interfaz de usuario común a los tres buscadores, y se ha puesto en funcionamiento sobre una colección de textos de prueba en inglés, correspondientes a las colecciones de noticias periodísticas extraídas del periódico Times. Esta interfaz se puede consultar a partir de la página principal del proyecto.
2. Se ha integrado dentro de la interfaz común un “localizador geográfico” que permite el acceso en lenguaje natural a una base de datos con información geográfica de la Universidad de Alicante.
3. Se ha desarrollado un sistema de desambiguación de sentidos, a partir de la base de datos léxica WordNet
4. Se ha desarrollado un sistema de análisis sintáctico parcial, que permite seleccionar constituyentes complejos en textos no restringidos, tanto en inglés como en español. Sobre este analizador sintáctico también funciona un sistema de resolución de anáfora, tanto de resolución de pronombres como de sintagmas nominales definidos y acrónimos.
5. Se ha desarrollado un módulo específico para el reconocimiento de entidades con nombre.
6. Como etiquetadores léxico-morfológicos se ha seleccionado el Tree-Tagger para inglés, y el RELAX para español. Como bases de datos de información semántica (sinónimos, hiperónimos, etc.) se está utilizando WordNet y EuroWordNet español.
7. En cuanto a los recursos lingüísticos, además de la colección extraída del Times, se está trabajando sobre las colecciones de los concursos CoNLL, TREC y CLEF, las cuales se encuentran disponibles en los idiomas que se pretenden tratar en este proyecto.

Finalmente, queda reseñar que parte del trabajo realizado ha sido divulgado mediante diferentes publicaciones y participaciones en los más importantes foros de investigación sobre recuperación de información, como por ejemplo las ediciones del año 2002 del TREC y CLEF multilingüe.