LLajú: Un sistema de recuperación multilingüe basado en EuroWordNet

Fernando Martínez Santiago
Manuel Carlos Díaz Galiano
L. Alfonso Ureña López
Maite Martín Valdivia
Manuel García Vega
José Ramón Balsas Almagro
Departamento de Informática. Universidad de Jaén. Spain
{dofer, mcdiaz, laurena, maite, mgarcia, jrbalsas}@ujaen.es

Resumen Se presenta aquí un sistema de recuperación de información multilingüe completamente funcional, capaz de procesar consultas en inglés y español, recuperando indistintamente documentos en ambos idiomas.

1 Introducción

El sistema de recuperación de información multilingüe LLajú aquí presentado viene a ser una implementación del modelo sugerido en [1], con la mejora en el cálculo de probabilidades de traducción a partir de SemCor tal como se describe en [2]. El sistema de recuperación que proponemos es capaz de recuperar artículos de las secciones de nacional e internacional de los sitios de "ABC", "El Mundo" y "El País", y de las secciones de internacional del "Washington Post", "CNN news" y "The Guardian Observer", correspondientes al año 2001. El proceso de recopilación y formateado del corpus se ha llevado a cabo usando la herramienta WebReader[3].

2 Descripción de Llajú

LLajú consta de tres partes bien diferenciadas:

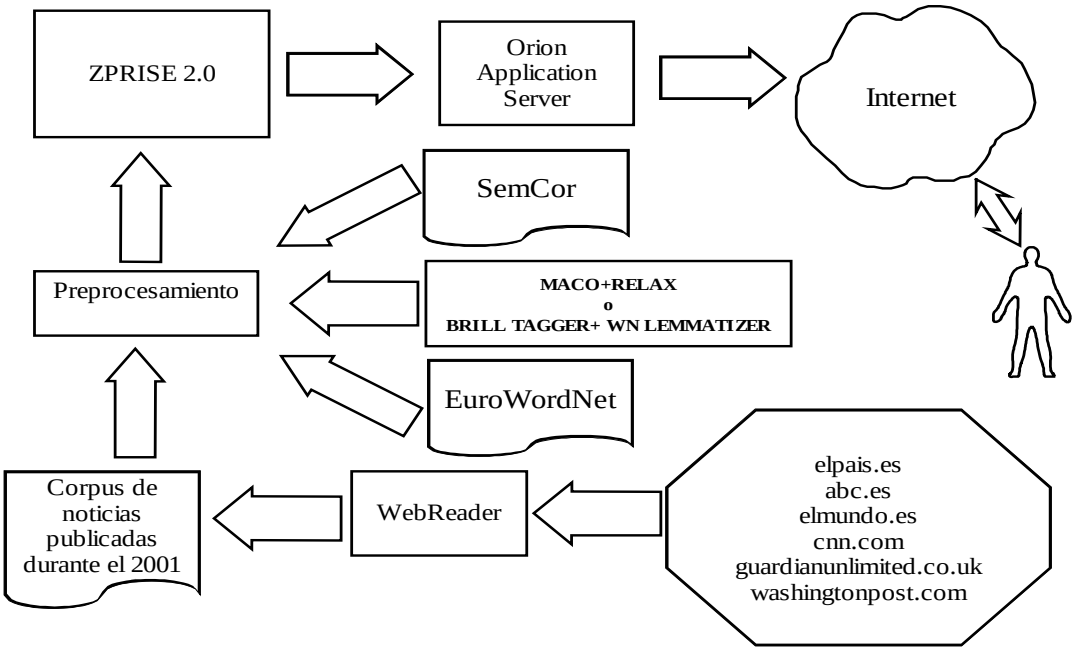
i. un motor de búsqueda multilingüe
ii. una interfaz basada en el Web
iii. una utilidad para la adquisición de nuevos documentos

El motor de búsqueda

El motor de búsqueda requiere un preprocesamiento, que pretende resolver las barreras lingüísticas existentes, para posteriormente indexar los documentos con independencia del lenguaje origen, usando para ello el sistema IR ZPrise. La elección de ZPrise ha venido determinada por su disponibilidad y por tratarse de un sistema recomendado en tareas CLIR como la que aquí se presenta. En cualquier caso, antes de indexar un documento, se requiere cierto preprocesamiento que a continuación describimos:

i. Detección de multi-palabras registradas en EuroWordNet.
ii. Extracción del lema para cada término que no forme parte de una multipalabra. Si se trata de un documento en español se ha usado MACO+RELAX [4]. En caso de tratarse de un término inglés, se ha optado por el Brill Tagger, junto con el lematizador que se encuentra en WordNet.
iii. Resolución de la ambigüedad léxica. Para esta tarea, se ha usado el desambiguador propuesto en [5].
iv. Obtención del *synset* al que pertenece el término ya desambiguado. De esta manera, conseguimos dos ventajas: por una parte estamos indexando por concepto antes que por términos, y además conseguimos un índice independiente de lenguaje.

Ilustración 1 - arquitectura del sistema LLajú



La interfaz Web

El sistema es accesible desde el Web. Para ello se ha optado por desarrollar un sistema basado en la tecnología Java 2 Enterprise Edition[1]. Concretamente se ha usado en servidor de aplicaciones Orion Application Server[2]. La arquitectura desarrollada permite aislar limpiamente las distintas capas de que consta el sistema en su totalidad (fig. 1), de tal forma que la interfaz Web mantiene un elevado grado de independencia respecto del sistema IR implementado.

WebReader

WebReader es una utilidad que permite crear corpus a partir del Web. Mediante la especificación de reglas y alguna heurística, es posible recuperar páginas de muy variados orígenes y formatos, para posteriormente obtener a partir de ellas un conjunto de documentos con un formato homogéneo, pero respetando el contenido original, permitiendo de esta manera su posterior procesamiento automático con fines experimentales en PLN o, más concretamente en nuestro caso, recuperación de información.

3 Trabajo futuro

La finalidad última de este proyecto es la consecución de una arquitectura adecuada sobre la cual se puedan probar los experimentos en tareas CLIR. Así, el fututo de Llajú pasa por mejoras en cuanto al manejo de multi-palabras y la resolución de la ambigüedad, así como la generalización del sistema a otros idiomas de la comunidad europea.

[1] En http://java.sun.com/j2ee

[2] En http://www.orionserver.com

4 Referencias

[1] Gonzalo et al. (1998). "Applying EuroWordNet to cross-language text retrieval". En *Computers and the Humanities*, 32(2-3), pp 185-207

[2] F. Martínez et al. “Evaluating SemCor in Cross-Language Information Retrieval Tasks“. En *Cross-Language Information Retrieval and Evaluation: Proceedings of the Second Cross-Language Evaluation Forum.* Pendiente de publicación.

[3] F. Martínez et al. “WWW como fuente de recursos lingüísticos para su uso en PLN”. En *XVII Congreso de la SEPLN*. En este mismo volumen.

[4] S. Acebo et al. (1994) “E*MACO: Morphological Analyzer Corpus-Oriented”*. En *ESPRIT BRA-7315 Acquilex II*, Working Paper 31.

[5] L.A. Ureña et al. (2001). “Integrating linguistic resources in TC through WSD. En *Computers and the Humanities*, 35/2, pp. 215-230. May

2001.