Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 33-40 recibido 1-05-2009; aceptado 5-06-2009

Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia*

Geo-NER: an English geographic entity recognizer based on GeoNames and Wikipedia

José Manuel Perea Ortega **Arturo Montejo Ráez**
Fernando Martínez Santiago **L. Alfonso Ureña López**

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{jmperea,amontejo,dofer,laurena}@ujaen.es

Resumen: En este artículo se presenta una herramienta para detectar y reconocer entidades específicamente geográficas. Esta herramienta se basa en la utilización de recursos externos tales como *gazetteers*, Wikipedia y reconocedores de entidades genéricos. Durante su funcionamiento se hace uso de cierto razonamiento geográfico, apoyándose en patrones sintácticos y características geográficas (términos tales como *río*, *montaña*, *lago*, etc.). Para su evaluación se han utilizado las consultas geográficas desarrolladas para la tarea GeoCLEF, enmarcada en las conferencias CLEF, obteniendo prometedores resultados para el reconocimiento de entidades de tipo geográfico, en comparación con otros reconocedores de entidades genéricos.
Palabras clave: Reconocimiento de entidades nombradas, recuperación de información geográfica

Abstract: In this paper we show a tool for detecting and recognizing geographic entities specifically. This tool is based on the use of external resources such as gazetteers, Wikipedia and generic entities recognizers. During its operation, it makes use of geographical reasoning, based on syntactic patterns and geographic features (terms such as *river*, *mountain*, *lake*, etc.). The evaluation was carried on the geographic queries developed for the task GeoCLEF, part of the CLEF conferences, obtaining promising results in the recognition of geographic entities, compared to other generic entity recognizers.
Keywords: Named entities recognition, geographic information retrieval

1. Introducción

Hoy día, la información geográfica se encuentra presente en una amplia variedad de medios y tipos de documentos. Durante las pasadas décadas, la tecnología empleada para acceder a este tipo de información se ha centrado en la combinación de mapas digitales y bases de datos, caracterizando a la mayoría de los Sistemas de Información Geográfica (*Geographic Information Systems*, GIS) (Chang, 2007; Bolstad, 2005). Sin embargo, en los últimos años se ha prestado especial atención al desarrollo de sistemas automáticos que recuperen información específicamente geográfica presente en documentos de texto no estructurados como los que componen la Web (Larson, 1996; McCurley, 2001; Jones y Purves, 2008).

La recuperación de información geográfica (*Geographic Information Retrieval*, GIR) puede considerarse una extensión de la recuperación de información tradicional (*Information Retrieval*, IR), incluyendo todas las áreas que tradicionalmente forman el núcleo de investigación en IR y haciendo un especial énfasis en el indexado y recuperación espacial y geográfica (Larson, 1996). Se ha comprobado que una de las partes fundamentales en una arquitectura GIR es el motor de recuperación de información utilizado (Perea-Ortega et al., 2008a). Un análisis general sobre los sistemas GIR presentados a la tarea

* Esta investigación ha sido parcialmente financiada por el Gobierno Español, proyecto TIMMOM (TIN2006-15265-C06-03), por la Universidad de Jaén, proyecto RFC/PP208/UJA-08-16-14 y por la Junta de Andalucía, Consejería de Turismo y Deporte (FFIEXP06-TU2301-2007/000024)

