Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 107-108

Aproximación a una estación lexicológica orientada a Internet

Zenón J. Hernández Figueroa
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
zhernandezfscarrera@dis.ulpgc.es

Director: Octavio Santana Suárez
Universidad de Las Palmas de Gran Canaria
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
osantana@dis.ulpgc.es

Resumen: Esta tesis es una proyección natural de los trabajos realizados por el Grupo de Estructuras de Datos y Lingüística Computacional de la ULPGC en los últimos años. Estos trabajos se han desarrollado en el ámbito de la Lingüística Computacional y han dado lugar, entre otros resultados, al desarrollo de herramientas de reconocimiento y generación morfológica. En esta tesis se propone la utilización de dichas herramientas como parte de nuevas aplicaciones cuyo objetivo es obtener provecho del enorme caudal de información lingüística que supone Internet. Se caracterizan dos clases de aplicaciones —en función del grado de interactividad de los estudios lingüísticos que se pretenda realizar— y se desarrollan sendos prototipos —denominados DAWeb y NAWeb— con una arquitectura estudiada para obtener los rendimientos más adecuados a cada caso. Las modalidades de análisis abarcan: la detección de neologismos, estudio del uso de las palabras con diversas medidas cuantitativas y cualitativas, y aspectos cercanos a la sintaxis tales como colocaciones léxicas o regímenes preposicionales.
Palabras clave: morfología, análisis de textos, Internet, lingüística computacional.

Abstract: This thesis follows up the works performed in the recent last years by the Data Structures and Computational Linguistics Group at ULPGC. These works have been developed about computational Linguistics and, as one of their results, some tools for morphologic identification and generation have been released. This thesis proposes the use of those tools as parts of new applications designed to benefit from the great linguistic information flow from Internet. Two kinds os applications are identified, both according to the interactivity of the linguistics studies to be made, and two prototypes, named DAWeb and NAWeb, are developed with special atention on their architecture in order to maximize the efficiency of both. Analysis modes include: neologism detection, word use (qualitative and quantitative measurements) and nearing sintaxt aspects like lexical colocations or prepositional regimes.
Keywords: morphology, text analysis, internet, computational linguistic.

Esta tesis es una proyección natural de los trabajos realizados por el Grupo de Estructuras de Datos y Lingüística computacional de la Universidad de Las Palmas de Gran Canaria en los últimos años. Estos trabajos se han centrado en el ámbito de la lingüística computacional y han dado lugar, entre otros resultados, al desarrollo de herramientas de reconocimiento y gestión morfológica, algunas de las cuales se encuentran disponible para su utilización en línea en la página web del grupo (http://gedlc.ulpgc.es). En esta tesis se propone la utilización de dichas herramientas como parte de nuevas aplicaciones cuyo objetivo es obtener provecho del enorme caudal de información lingüística que supone Internet.

Tras realizar el preceptivo análisis de requerimientos, se caracterizan dos clases de aplicaciones —en función del grado de interactividad de los estudios lingüísticos que se pretenda realizar— y se desarrollan en consecuencia sendos proptotipos informáticos —denominados DAWeb y NAWeb— con una arquitectura estudiada para obtener los rendimientos más adecuados a cada caso.

NAWeb está diseñado para la exploración en detalle de documentos individuales bajo supervisión directa del usuario y reúne las

ISSN 1135-5948

características típicas de un navegador web pero además incorpora una amplia variedad de opciones para el análisis de las páginas accedidas.

DAWeb se orienta al estudio conjunto de grandes vólumenes de documentos de forma desasistida y adopta el formato de un descargador de páginas con la diferencia de que en vez de bajar las páginas que accede, las analiza y almacena sólo los resultados.

Las modalidades de análisis que pueden realizar ambos prototipos abarcan: (1) en primer lugar la detección de neologismos, entendiendo como tal en primera instancia cualquier palabra que las herrmientas de reconocimiento morfológico no identifiquen —luego hay que filtrar si se trata de entidades tales como nombres propios, secuencias especiales o incluso simples errores ortográficos—, (2) en segundo lugar, realiza diversas medidas cuantitativas y cualitativas para el estudio del uso de las palabras, y (3) finalmente, trata con aspectos cercanos a la sintaxis tales como el estudio de colocaciones léxicas o regímenes preposicionales.

Las dos aplicaciones comparten los módulos relacionados con la obtención de los textos y su posterior análisis; ambas incluyen un módulo optimizador de búsqueda morfológica que aumenta sustancialmente la velocidad de reconocimiento gracias a la utilización de una estructura de datos de acceso rápido.

DAWeb se caracteriza por tener una arquitectura orientada al procesamiento en paralelo para minimizar los costos de acceso a Internet, mientras que el aspecto más destacado de NAWeb es su marcada interactividad.

Ambas aplicaciones aportan un novedoso complemento al concepto de estación lexicológica o estación filológica que algunos —especialmente en el campo de la lexicografía— han postulado con anterioridad; los autores centraban su atención principalmente en la gestión de la información disponible —mediante sistemas de bases de datos— y la generación de productos a partir de la misma —diccionarios en el caso de los lexicógrafos. Se trata de un concepto integrador que plantea la reunión sobre una misma plataforma tecnológica integrada de los textos objeto de estudio y todas las herramientas necesarias para su manipulación.

El mayor escollo con el que se ha encontrado en el pasado el uso de este tipo de estaciones, y en general cualquier planteamiento tendente a la utilización de ordenadores para el estudio del lenguaje, radicaba en la escasez de documentos en formato electrónico, lo que obligaba al estudioso a formar trabajosamente su propio corpus de referencia —distrayendo esfuerzos de su verdadero objetivo—, o bien a conformarse con lo que estuviera disponible —que podía ser incompleto y quizás poco adecuado. El explosivo desarrollo de Internet y su previsible continuidad alivian esta carencia siempre que se disponga de herramientas adecuadas para sacarle partido.

En este sentido, resulta interesante lo aportado en el Anuario 2000 de la serie “El español en el mundo”, que elabora el Instituto Cervantes , donde se afirma que “...la presencia firme del español en Internet exige asegurar para el español una banda de uso de la red de redes de entre un 15 y un 25 por ciento en los próximos cuatro años...” y que “... El aumento de la presencia del español sólo puede lograrse mediante un incremento de los contenidos en español y un desarrollo equiparable de los sistemas de recuperación de información y su análisis...”.

En esta tesis el foco se pone en la fase de obtención de información lingüística a partir de una fuente como la metarred, no disponible en el pasado, pero con una fuerte proyección de futuro. Constituye un importante instrumento enmarcado dentro de lo que el Instituto Cervantes llama “...desarrollo equiparable de los sistemas de recuperación de información y su análisis...”

Además de sus funciones principales, para las que han sido diseñadas, las aplicaciones desarrolladas pueden —especialmente NAWeb— tener otras utilidades como el estudio de estilos o la enseñanza del español a extranjeros; con este fin se ha dotado a NAWeb de la capacidad de analizar otros formatos —texto plano, documentos de Microsoft Word— distintos del típico HTML de las páginas Web.

Los prototipos se han desarrollado en el entorno Microsoft Windows combinando los lenguajes Delphi (ObjectPascal + herramientas RAD) para el desarrollo de interfaces de usuario y sistemas de navegación y descarga, y C++ para las herramientas de reconocimiento y desambiguación morfológica.