ITEM: Recuperación de Información Textual en un Entorno Multilíngüe con Técnicas de Lenguaje Natural.

Organismo Financiador: Comisión Interministerial de Ciencia y Tecnología

Grupos Participantes en el Proyecto:

- Grupo UNED en Procesamiento de Lenguaje Natural
- Grupo UPC en Procesamiento de Lenguaje Natural
- Grupo UPV/EHU en Procesamiento de Lenguaje Natural
- Grupo EB en Procesamiento de Lenguaje Natural

Personas de contacto:

- Dra. Mª Felisa Verdejo. Coordinadora del Proyecto. Dept. IEEC. ETSI Industriales UNED Madrid. Email: felisa@ieec.uned.es.
- Dr. Horacio Rodríguez. Investigador responsable grupo UPC. Departamento de Lenguajes y Sistemas Informáticos. Universidad Politécnica de Cataluña. Email:horacio@lsi.upc.es
- Dra. Arantza Díaz de Ilarraza. Investigadora responsable grupo UPV/EHU. Departamento de Lenguajes y Sistemas Informáticos. Facultad de informática. Universidad del País Vasco. Email: jipdisaa@si.ehu.es

Dirección de contacto: http://sensei.ieec.uned.es/item

Resumen

El objetivo principal del proyecto es explorar y evaluar en qué medida el uso de técnicas de procesamiento del lenguaje natural puede mejorar los procesos de recuperación y de extracción de información en sistemas de datos textuales o multimediales.

El proyecto explora las técnicas de procesamiento de lenguaje natural aplicadas a la recuperación de información en dos vertientes: Por un lado, en la incorporación de tecnologías lingüísticas a los sistemas clásicos de recuperación de información (básicamente estadísticos), para mejorar tanto la indiciación de textos como su posterior consulta. Por otro lado, en aumentar el número de usuarios potenciales, facilitando el acceso mediante un sistema de consulta multilingüe a las bases de datos documentales.

Para ello, un primer paso es la implantación e integración de diversas herramientas lingüísticas ya disponibles, en una plataforma común estandar. A continuación, el proyecto plantea el desarrollo de técnicas de parsing robusto -según las necesidades de consulta e indiciación- y la creación de una base de conocimiento conceptual común para todas las lenguas contempladas.

El proyecto está organizado en 7 tareas que se describen a continuación:

T0-Coordinación
Duración: todo el proyecto
Estado: activa

T1-Implantación e integración de herramientas en una plataforma estandar.
Duración: 24 meses
Estado: activa

T2-Investigación en técnicas robustas de análisis de texto libre en lenguaje natural.
Duración: 24 meses
Estado:activa

T3-Bases de conocimiento ontológicas
Duración: 24 meses
Estado: activa

T4-Base documental textual y prototipo de interfaz
Duración: 12 meses
Estado: activa

T5-Extracción de información
Duración: 18 meses
Estado: activa

T6-Implantación, prueba y evaluación
Duración: 12 meses
Estado: sin inicializar.