Procesamiento del lenguaje Natural, nº 41 (2008), pp. 317-318 recibido 7-05-2008; aceptado 16-06-2008

TEXT-MESS: Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano*

TEXT-MESS: Intelligent, Interactive and Multilingual Text Mining based on Human Language Technologies

Patricio Martínez-Barco
Manuel Palomar
Univ. de Alicante
GPLSI - DLSI
patricio@dlsi.ua.es
mpalomar@dlsi.ua.es

Julio Gonzalo
Anselmo Peñas
UNED
GPLN - DLSI
julio@lsi.uned.es
anselmo@lsi.uned.es

L. Alfonso Ureña-López
Mª Teresa Martín-Valdivia
Univ. de Jaén
SINAI - DI
laurena@ujaen.es
maite@ujaen.es

Ferran Pla
Paolo Rosso
Univ. Pol. de Valencia
GPLIS,GRFIA - DSIC
fpla@dsic.upv.es
prosso@dsic.upv.es

Alicia Ageno
Jordi Turmo
Univ. Pol. de Catalunya
GPLN,TALP - DLSI
ageno@lsi.upc.edu
turmo@lsi.upc.edu

M. Antònia Martí
Mariona Taulé
Univ. de Barcelona
CLIC
amarti@ub.edu.es
mtaule@ub.edu.es

Resumen: El objeto de este proyecto es analizar, experimentar y desarrollar tecnologías inteligentes, interactivas y multilingües de minería de textos, como pieza clave de la próxima generación de motores de búsqueda y análisis textual, sistemas capaces de encontrar "la necesidad que subyace a la consulta". Estas tecnologías ofrecerán servicios e interfaces especializadas según el dominio y el tipo de necesidad de información. Además, integrarán búsqueda documental (páginas web), multimedia (imágenes, audio, video), en información semiestructurada y en dominios específicos.
Palabras clave: Minería de textos, Tecnologías del Lenguaje Humano (TLH), Recursos de TLH, Recuperación de Información, Búsqueda de Respuestas, Extracción de Información, Evaluación de TLH, CICYT

Abstract: The goal of this project is to analyze, experiment, and develop intelligent, interactive and multilingual Text Mining technologies, as a key element of the next generation of search engines, systems with the capacity to find "the need behind the query". These technologies will provide specialized services and interfaces according to the search domain and type of information needed. Moreover, it will integrate searchs on document collections (websites), multimedia (images, audio, video), semi-structured texts and restricted domains.
Keywords: Text Mining, Human Language Technologies (HLT), HLT resources, Information Retrieval, Question Answering, Information Extraction, HLT Evaluation, CICYT

1. Descripción general

La gran cantidad de información disponible actualmente en formato electrónico junto al creciente número de usuarios finales que disponen de acceso directo a dicha información a través de ordenadores personales, ha impulsado la investigación en sistemas de información textual que faciliten el análisis, la localización, la gestión, el acceso y el tratamiento automático de toda esta ingente cantidad de datos.

Internet ha cambiado profundamente la forma en la que las personas se comunican, negocian y realizan el trabajo diario, al tener acceso a infinidad de recursos, en diferentes formatos y en diferentes idiomas. Todos estos factores han contribuido al éxito de la Web y a la vez ha originado, paradójicamente, uno de sus principales problemas el exceso de información.

En este marco de sobrecarga de información los actuales motores de búsqueda

* TIN2006-15265-C06

han quedado obsoletos. Por este motivo, las seis universidades integrantes del proyecto TEXT-MESS trabajan bajo el patrocinio del actual Ministerio de Ciencia e Innovación con el fin de definir una nueva generación de motores de búsqueda capaces de encontrar "la necesidad detrás de cada consulta" y que ofrecerán servicios e interfaces especializadas según el dominio y el tipo de necesidad de información. Además integrarán búsqueda documental (paginas web), búsqueda multimedia (imágenes, audio, video) y búsqueda en bases de datos (biomedicina, turismo, bolsas de empleo, etc.). Los nuevos buscadores serán capaces de descubrir y organizar la información, y no sólo de producir listas ordenadas de páginas web.

En estos nuevos buscadores las tecnologías del lenguaje jugarán un papel más relevante que en los motores de búsqueda actuales que además han venido dando prioridad a los contenidos en inglés y es tecnología estadounidense en una gran proporción. Así TEXT-MESS cumple la doble misión estratégica de definir el papel de las tecnologías del lenguaje en estos nuevos sistemas y de posicionar las lenguas oficiales del estado en esa "carrera" tecnológica que actualmente está ya lanzada.

2. Objetivos

La finalidad del proyecto es desarrollar tecnologías inteligentes, interactivas y multilingües de minería de textos, que integren la búsqueda documental en páginas web, la búsqueda multimedia sobre imágenes y la búsqueda sobre información semiestructurada, que se basen en TLH.

Para llevar a cabo este objetivo, se proponen tres líneas de actuación.

(1) Desarrollar sistemas de minería de textos (búsqueda, extracción, análisis, clasificación y recuperación de información), estudiando por un lado los aspectos multilingües (con especial énfasis en el español y catalán) e interactivos, la eficacia y eficiencia de los sistemas sobre documentos escritos, transcripciones de audio e imágenes, trabajando además tanto en dominios genéricos (la Web) como específicos (como es el caso de la biomedicina y turismo). (2) Mejorar y adaptar los recursos y las herramientas existentes (mayor cobertura, calidad y tratamiento de dominios específicos) y crear nuevos recursos, técnicas y herramientas necesarias para abordar las nuevas aplicaciones basadas en Tecnologías del Lenguaje Humano combinando conocimiento lingüístico y técnicas de aprendizaje automático (machine learning). (3) Entroncar el proyecto con las principales campañas internacionales de evaluación de sistemas de búsqueda y Tecnologías del Lenguaje Humano; por un lado, como participantes en estas campañas, para contrastar los resultados de nuestra investigación con los mejores grupos de investigación a nivel internacional; por otro lado, como promotores y coordinadores de algunas tareas, con el objetivo de promover la investigación en las líneas de interés de este proyecto y garantizar la presencia, en condiciones de igualdad, de los idiomas de interés del proyecto (español y catalán) en la investigación competitiva en este campo.

Para la consecución del objetivo global y el desarrollo óptimo de las diferentes líneas de actuación del proyecto descritas previamente, éste se ha propuesto a través de un proyecto coordinado que consta de los subproyectos:

- **01-KRUA** Knowledge discovery and Representation in Human Language Technology (UA);
- **02-INES** Intelligent exploration and synthesis of search results (UNED);
- **03-TIMOM** Tratamiento de Información multiMOdal y Multilingüe (UJA);
- **04-MiDEs** Métodos de Aprendizaje para Minería de Textos en Dominios Específicos (UPV);
- **05-SAMiT** Sistemas Adaptativos de Minería de Textos (Text Mining Adaptative Systems) (UPC);
- **06-Lang2World** Discovering world knowledge coded into language (UB).

3. Estado actual

TEXT-MESS tiene una duración de tres años (octubre 2006 - septiembre 2009) y actualmente ya dispone de resultados tanto de investigación (con más de un centenar de publicaciones en congresos y revistas de prestigio de ámbito internacional) como de recursos generados (corpus, herramientas, técnicas), así como prototipos de aplicación de la investigación que pueden encontrarse en la página web del proyecto[1].

[1] http://gplsi.dlsi.ua.es/text-mess