GRUPO COLE

universidade
SUPLEMENTO DE EDUCACIÓN DE GALICIA HOXE
Número 27 / 2009 · 13 de outubro de 2009

Investigadores da UDC e da Uvigo colaboran neste singular proxecto

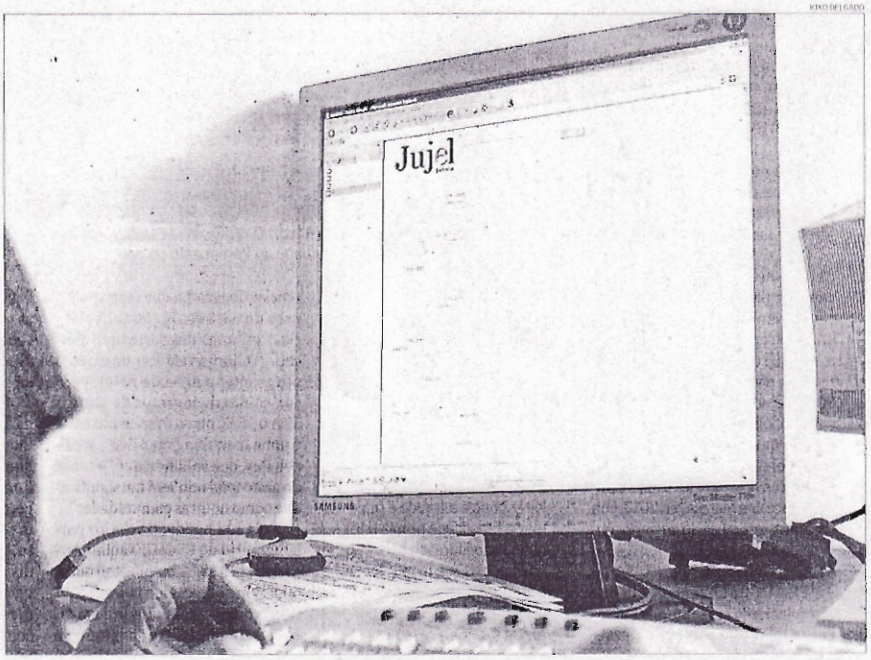
OS ORDENADORES TAMÉN PODERÁN LER EN GALEGO

S.U. . SANTIAGO

Varios investigadores do Grupo COLE (Compiladores e Linguaxes), da Escola Superior de Enxeñaría Informática da Universidade de Vigo, dirixido polo catedrático Manuel Vilares Ferro, levan xa varios anos inmersos na investigación da recuperación de información.

Tanto é así, que Juan Otero Pombo, un dos investigadores, presentou recentemente a súa tese de doutoramento sobre este tema, baixo o título *Análisis léxico robusto*, na cal aborda o desenvolvemento e a avaliación de técnicas de corrección ortográfica e a súa aplicación en contornos de recuperación de información nos que os erros ortográficos están presentes. Trataríase de desenvolver un sistema que permita identificar correctamente as palabras con erros ortográficos que un usuario introduce nunha busca, por exemplo en Google ou na base de datos dunha biblioteca.

Tal como se recolle na tese de Juan Otero Pombo, hoxe en día hai unha "frenética e constante evolución da informática" que deu lugar ó que se coñece como Sociedade da Información. Na súa investigación, Otero Pombo, que estivo seis anos recabando datos, asegura que actualmente xéranse e publícanse numerosos datos en formato electrónico, que deben ser procesados e estruturados para facilitar o acceso rápido e eficaz. Porén, "a maior parte dos contidos atópase con escaso ou, simplemente, sen ningún tipo de estrutura". É aquí onde entran os sistemas de recuperación de información, que permiten localizar aqueles documentos dunha colección que satisfan os requirimentos dun usuario.



Funcións como as procuras en internet simplificaríanse cun software que lle permitise ó ordenador entende ro que queremos

PRETÉNDESE DOTAR O NOSO IDIOMA DUN NOVO PROGRAMA INFORMÁTICO DE PROCESAMENTO DA LINGUAXE NATURAL

Polo tanto, tal como defende Otero Pombo, faise necesario contar con mecanismos eficaces que, desde o punto de vista computacional, permitan que as persoas poidan comunicarse directamente coas máquinas en linguaxe humana, e que estas sexan capaces de extraer coñecemento a través da linguaxe natural.

Por iso xorde o Procesamento da Linguaxe Natural (PLN), subdisciplina da Intelixencia Artificial que persegue a representación e comprensión automática da linguaxe humana por medio da realización dunha análise léxica (recoñece unidades mínimas da linguaxe), sintáctica (recoñece unidades gramaticais formadas por varias unidades léxicas), semántica (captura o significado dunha frase) e pragmática (engádelle información ó significado da frase en función do contexto). Así e todo, na busca de información, como actividade humana que é, sucede frecuentemente a "introducción de erros ortográficos ou de dixitación na consulta, o cal complica a tarefa de recuperación".

Por todo isto, o obxectivo é desenvolver e avaliar a tecnoloxía de base necesaria para o PLN, sobre todo no ámbito da análise léxica e da corrección ortográfica e a etiquetación. Así, tras estas investigacións, Juan Otero Pombo aportou coa súa tese un novo método de corrección ortográfica "máis eficiente e menos custoso que os que se viñeron aplicando ata o momento, xa que reduce ó mínimo posible a exploración do dicionario en busca da corrección máis adecuada en cada momento". Esta técnica integrouse despois no etiquetador morfosintáctico MrTagoo, desenvolvido polo grupo COLE da Universidade de Vigo e o LyS (Lengua y Sociedad de la Información) da Universidade da Coruña, que permite asignar automaticamente unha categoría léxica (substantivo, verbo, ...) a cada palabra dun texto en español ou galego.

O resultado desta integración permite solucionar os tres problemas comúns no proceso de consultas nun sistema computacional de recuperación de información: a ambigüidade segmental (cando a identificación dos termos que forman unha consulta non resultan triviais), a ambigüidade morfosintáctica (cando as palabras poden xogar distintos papeis segundo a frase na que aparezan) e a corrección ortográfica contextual (elixe entre as alternativas de corrección a que mellor encaixa coa consulta). O resultado dos experimentos, realizados nunha contorna de recuperación de información con consultas degradadas, poñen de manifesto que "o emprego de técnicas de corrección ortográfica ten un impacto moi positivo sobre os sistemas de recuperación de información", fronte a outras propostas realizadas con anterioridade.

O OBXECTIVO É FACILITAR A INTERACCIÓN DE HUMANOS E COMPUTADORAS

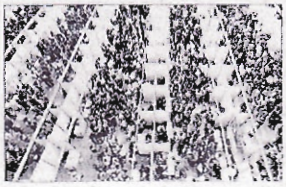
No proxecto quíxose dotar dun PNL a lingua galega, e para iso naceu o proxecto *Victoria*, con equipos franceses e españois, nos cales se integraron Miguel Ángel Molinero (investigador do grupo LyS da Universidade da Coruña) e Elena Sánchez Trigo (investigadora do grupo COLE da Universidade de Vigo). Nunha primeira fase, os investigadores de *Victoria* concentráronse nos recursos necesarios para construír analizadores sintácticos para español e galego.

Actualmente, o proxecto xa construíu varios recursos que serán liberados baixo licenza LGPL-LR (Lesser General Public License for Linguistic Resources), entre os cales destacan regras de configuración idiomáticas e un léxico con información morfolóxica para a lingua galega. ●

Carlos Pajares, delegado do CERN en España, inaugurou o Máster de Física Nuclear da USC

O catedrático da universidade compostelá, delegado no noso país da Organización Europea para as Investigacións Nucleares, abriu cunha conferencia este novo ciclo formativo. **Páxina 12**

A Universidade da Coruña preparou para finais de mes a II Feira Europea do Emprego



A institución herculina, co apoio da Deputación da Coruña e da Xunta de Galicia organizou unha nova edición deste foro no cal se poden reunir e coñecer empresas e estudantes. **Páxina 13**