El sistema ERIAL: LEIRA, un entorno para RI basado en PLN*

Fco. Mario Barcala y **Eva M.ª Domínguez**
Centro Ramón Piñeiro para a Investigación en Humanidades

M. A. Alonso y **D. Cabrero** y **J. Graña** y **J. Vilares** y **M. Vilares**
Universidade da Coruña

Guillermo Rojo y **M.ª Paula Santalla** y **Susana Sotelo**
Universidade de Santiago de Compostela

Resumen En la demostración aquí descrita se expone el sistema ERIAL para RI basada en PLN. La demostración presenta el entorno LEIRA desarrollado con tal fin en el marco del proyecto ERIAL. Para los propósitos de la demostración se ha indexado un corpus cedido por Editorial Compostela de 20.000 textos de noticias de prensa en gallego y español. La interfaz se ha adaptado para permitir una serie de interrogaciones que han sido previamente contrastadas a mano con el corpus, pero continúa, asimismo, permitiendo la formulación de cualquier interrogación por parte del usuario.

1 Introducción: qué, quién, cómo

¿Qué? LEIRA es un entorno para recuperación de información basado en PLN y desarrollado en el marco del proyecto ERIAL, *Extracción y Recuperación de Información mediante Análisis Lingüístico*, un proyecto de investigación que aspira a mejorar los resultados de las aplicaciones de RI en gallego y español mediante el uso de técnicas de PLN.

¿Quién? ERIAL ha sido desarrollado por un grupo de investigación interdisciplinar integrado por la Universidad de A Coruña, que aportaba el conocimiento computacional, las Universidades de Santiago de Compostela y Vigo y el Centro Ramón Piñeiro para la Investigación en Humanidades, que aportaban conocimiento lingüístico, respectivamente, sobre español y gallego, la empresa Compaq, que aportaba equipamiento y asesoramiento técnico, y la empresa Editorial Compostela, como usuario final del sistema. El proyecto fue financiado, entre los años 1999 y 2001, por la Secretaría de Estado de Política Científica y Tecnológica, con Fondos Europeos de Desarrollo Regional (FEDER).

¿Cómo? Los términos de indexación extraídos para cada texto sobre el que se hayan de lanzar las búsquedas, así como los extraídos de la interrogación formulada por el usuario y que hayan de ser comparados con aquéllos, codifican relaciones lingüísticas consistentes en i) lematización, ii) derivación morfológica, iii) sintaxis. Las interrogaciones, simples o complejas, pueden formularse en lenguaje natural.

2 Descripción y presentación de la interfaz LEIRA

La interfaz actual permite los siguientes cinco tipos diferentes de búsquedas, cada uno de los cuales representa la explotación de un nivel diferente de procesamiento lingüístico tanto de los textos investigados como de la interrogación formulada:

- **L** Lematización
- **M** Expansión morfológica
- **S** Estructuras sintácticas
- **T** Co-ocurrencia total
- **P** Texto Plano[1]

El módulo que genera los términos sobre los que trabajan los indexadores se aplica a los textos una vez que éstos ya han experimentado un cierto procesamiento lingüístico:

* Véase Fco. M. Barcala et al., *Una aplicación de RI basada en PLN: el proyecto ERIAL*, en este mismo volumen.

[1] Este último tipo de búsqueda representa la aproximación tradicional a RI y, obviamente, ha sido sólo temporalmente incluido en la interfaz para, en el marco de esta demostración, permitir la comparación de resultados entre estas aproximaciones y las aquí propuestas.

los textos han sido primero pre-procesados, y después etiquetados y lematizados por el módulo etiquetador/lematizador. El módulo de indexación opera sobre los textos lematizados y puede aplicar asimismo ulteriores normalizaciones de los textos que codifican relaciones entre palabras que pertenecen a las mismas familias morfológicas, así como relaciones entre palabras de los textos que constituyen pares de dependencia sintáctica, identificados por medio de patrones de secuencias de etiquetas sobre los que luego se aplican reglas de extracción de pares de relación sintáctica directa. Para poder llevar a cabo las comparaciones requeridas durante la operación de recuperación de documentos en sí misma, las interrogaciones formuladas por los usuarios tienen que experimentar tanto los procesamientos lingüísticos previos mencionados arriba, como los más elaborados, requeridos por los más complejos tipos de indexación y búsqueda (familias morfológicas y pares de dependencia sintáctica).

Para los propósitos de la demostración aquí presentada, se ha indexado un corpus de 20.000 noticias de prensa, 10.000 en español y 10.000 en gallego, de unas 450 palabras de media por noticia, publicadas entre el 23 de noviembre de 1999 y el 10 de noviembre de 2000. Todos esos textos fueron leídos para i) identificar interrogaciones relevantes para series de documentos —se identificaron 230—, ii) identificar todos los documentos relevantes para las interrogaciones identificadas. Para los propósitos de la demo, la interfaz incluye una batería de interrogaciones aleatoriamente seleccionadas entre las 230 identificadas, así como, por supuesto, la posibilidad de formular cualquier interrogación solicitada por un usuario. Veamos más en detalle, ahora, en qué consiste cada tipo de búsqueda.

L–Lematización. Los términos indexados para cada texto, así como los extraídos de la interrogación que se compara con ellos, codifican relaciones que consisten en la pertenencia al mismo lema. Para ello, todos los nombres, adjetivos y verbos se reducen a sus lemas antes de llevar a cabo los procesos de indexación o comparación. Esta posibilidad de búsqueda se basa en nuestra aproximación en el recurso a un lexicón de formas de palabras cada una de ellas conectada con su lema.

M–Expansión morfológica. Los términos indexados para cada texto, así como los extraídos de la interrogación que se compara con ellos, codifican relaciones que consisten en la pertenencia a la misma familia morfológica, esto es, están, básicamente, relacionados por derivación. Con tal fin, los nombres, adjetivos y verbos son, no sólo reducidos a su lema, sino además a un representante canónico, arbitrariamente elegido, de la familia morfológica a la que pertenecen según un diccionario previamente generado a partir de todos los lemas del lexicón. Este proceso previo de generación de familias morfológicas da cuenta de una gran variedad de fenómenos morfológicos, especialmente de la sufijación y la prefijación derivativa, así como de alomorfias con ellas relacionadas.

S–Pares de dependencia sintáctica. Los términos indexados para cada texto, así como los extraídos de la interrogación que se compara con ellos, codifican relaciones que consisten en conexiones sintácticas tales como las establecidas entre núcleos y modificadores, verbos y objetos, etc. Para obtener esas relaciones, primero se identifican patrones sintácticos a partir de expresiones regulares que representan secuencias de etiquetas morfosintácticas. Una vez identificadas estas secuencias, reglas de extracción de pares de relación sintáctica directa extraen los pares de dependencia que constituirán los términos de indexación o comparación. Sobre los pares así obtenidos, opera aún la normalización morfológica descrita en el párrafo anterior.

T–Co-ocurrencia total. Es éste un parámetro distinto, no lingüísticamente motivado, que puede combinarse con cualquiera de los tipos de búsqueda previos. Se basa en la utilización de un motor de indexación booleano, en lugar de vectorial, como en los casos anteriores. Ello tiene como consecuencia que, para que un documento sea seleccionado, todos los términos presentes en la interrogación deben ser hallados también en el documento en cuestión. En la presente implementación, este parámetro se combina con la normalización de términos basada en la lematización.