LA LEMATIZACIÓN EN ESPAÑOL: UNA APLICACIÓN PARA LA RECUPERACIÓN DE INFORMACIÓN

Gómez Díaz, Raquel
Gijón: Ediciones Trea, 2005.
Colección Biblioteconomía y Administración Cultural; 125
ISBN: 84-9704-186-0

La recuperación de información (RI) es una disciplina relativamente joven, que ha experimentado un rápido desarrollo en los últimos años, debido fundamentalmente a la aparición de los motores de búsqueda. Sin embargo, desde su origen, se ha pronosticado que la lingüística computacional y las técnicas de procesamiento del lenguaje natural (PLN) traerían una auténtica revolución a los sistemas de recuperación de información (SRI). La obra que aquí se reseña aborda un aspecto indispensable en la investigación en este campo: la *lematización*, neologismo que se aplica al proceso de eliminación automática de partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (*lema*) y facilitar la eficacia de la indización y la consiguiente recuperación.

Estos neologismos, *lema* y *lematización*, proceden del campo de la Informática. En inglés se denominan *stem* y *stemming*, términos que encontraremos con frecuencia en la bibliografía internacional o en la propaganda de programas como dtSearch de Bitext. El *lema* es una etiqueta informática, que en español coincidirá generalmente con el *lexema* o *raíz* de las palabras, pero que no necesariamente han de ser equivalentes.

La autora es licenciada en Documentación por la Universidad de Salamanca y actualmente profesora de su Facultad de Traducción y Documentación. Este libro es en realidad una adaptación de su tesis, titulada «Estudio de la incidencia del conocimiento lingüístico en los sistemas de recuperación de la información para el español», leída en la Universidad de Salamanca en 2001. Los escasos cambios realizados al texto de la tesis no justifican el inexcusable retraso de su edición en formato de monografía. Sin embargo, a pesar del tiempo transcurrido, el tema sigue siendo de plena actualidad y de gran interés para los desarrollos innovadores en los sistemas de gestión de información

Como puede esperarse de una tesis doctoral, en la obra se puede encontrar un desarrollo teórico (definición de conceptos y valoración del estado de la investigación), junto con una aplicación práctica. En la síntesis teórica que aporta esta monografía se distinguen dos apartados. En primer lugar se aborda la definición de modelos de recuperación de información (RI) y los indicadores para su evaluación. En segundo lugar se centra en el concepto de *lematización* y la tipología de algoritmos que se aplican en este proceso. Finalmente, se presenta un trabajo práctico, que consiste en la elaboración de un *lematizador* específicamente diseñado para las características lingüísticas del castellano, y la evaluación de los resultados obtenidos en su aplicación en un SRI.

Un *lematizador* es un programa, basado en diferentes algoritmos, que trabaja sobre un corpus de textos en lenguaje natural y realiza una extracción automática de términos simplificados a su esencia o *lema*. También se aplica al análisis de las preguntas que se realizan al sistema en la fase de RI, se reducen los términos empleados por el usuario

antes de localizar las coincidencias en los índices. La autora compara los resultados obtenidos mediante diferentes técnicas: *lematización* flexiva (eliminación de plurales, género y desinencias verbales) frente a *lematización* derivativa (que elimina además sufijos derivativos), concluyendo que el primer modelo es el más eficaz en la relación precisión-exhaustividad, además de ser más sencillo de aplicar. También se concluye que la eliminación de palabras vacías contribuye a optimizar la recuperación.

La experimentación se realizó evaluando los resultados obtenidos sobre una base de datos referencial, Datathéke. Habría sido de mayor interés aplicar estas pruebas sobre una base de datos de documentos a texto completo. La autora lamenta la ausencia de colecciones experimentales en español. Este hecho es un síntoma de las carencias que sufrimos en España para la investigación, se carece de herramientas contrastadas de trabajo, los grupos de investigación no han desarrollado planes a largo plazo apoyados en el diseño de bases de datos experimentales, de gran utilidad para este tipo de evaluaciones.

La lematización puede entenderse como una fase en la construcción de un sistema de indización automática y/o de búsqueda experta. Sin embargo, en este libro se analizan sus resultados de forma aislada, sin tener en cuenta otras facetas del análisis lingüístico, necesarias para desambiguar los términos en la recuperación. Por ejemplo no se hace ninguna mención, en el desarrollo práctico, a la imprescindible diferenciación entre nombres propios y comunes, o a los términos que cambian de sentido según el contexto.

Las limitaciones de la experimentación que presenta este trabajo no le restan interés, pero es imprescindible que se vea continuada por más aplicaciones prácticas y por el análisis de cómo pueden insertarse estas herramientas en los actuales sistemas bibliográficos. La investigación desarrollada se enmarca entre tres disciplinas: Informática, Lingüística y Ciencias de la Documentación. Es importante que ésta última no se quede atrás y que los investigadores de nuestro ámbito se integren en equipos multidisciplinares. En la propia Universidad de Salamanca existe un grupo de investigación especializado en RI (http://reina.usal.es/), formado exclusivamente por informáticos. La principal aportación que puede hacerse desde las Ciencias de la Documentación radica en el enfoque pragmático, en la aplicación al desarrollo de nuevos modelos y recursos de información de utilidad real. La experiencia acumulada en las tareas documentales debería tomarse como base y no quedar al margen de la investigación en RI.

Luis Rodríguez Yunta
CSIC – CINDOC