La lematización en español: una aplicación para la recuperación de información

Raquel Gómez Díaz

EDICIONES TREA, S. L.

ÍNDICE

PRÓLOGO

La recuperación de información es un tema candente. No hay más que ver cómo aumenta el número de libros nuevos sobre el tema; cómo aparecen nuevas revistas dedicadas expresamente a la recuperación de información, o cómo incluso otras más genéricas cada vez incluyen más artículos sobre esta; cómo, en fin, los congresos, simposios, seminarios y similares se hacen ya imposibles de seguir, salvo que uno tuviera el don de la ubicuidad.

La causa es clara: además de la consabida explosión documental (ya se sabe, la cantidad de documentos crece, no sé si exponencialmente, pero en cualquier caso, muchísimo), fundamentalmente obedece a dos factores: la disponibilidad creciente de documentos en formato electrónico, y la formidable capacidad de difusión que proporciona Internet. Esa disponibilidad de documentos íntegramente en formato electrónico posibilita su procesamiento automático, con programas informáticos, y ello hace que podamos plantearnos la total automatización del proceso de recuperación.

Esta automatización supone, entre otras cosas, la superación de la indización manual; esto es, la construcción manual de representaciones más o menos formales del contenido de cada uno de los documentos. Supone también la superación de los lenguajes controlados y afines, utilizados para construir dichas representaciones; antes bien, la automatización del proceso tiene mucho que ver con el uso del lenguaje natural en sus dos vertientes: la del propio documento, por supuesto, y la del usuario cuando trata de expresar sus necesidades de información.

De manera que la recuperación de información contemporánea tiene poco que ver con tesauros, encabezamientos de materias o códigos de diversa especie; y mucho menos con la asignación manual de estos para describir el contenido de documentos que están en su integridad en formato electrónico y a los que se puede

acceder directamente a través de Internet. Si tuviéramos que esperar a que se realizase la indización manual de tales documentos para poder recuperarlos, no los encontraríamos nunca, por la sencilla razón de que la velocidad de aparición de documentos es mucho mayor que nuestra capacidad ideal de indizarlos manualmente. Esto, sin entrar en cosas como la inconsistencia de la indización manual, bien conocida, por otra parte, desde hace años.

Así que lo que necesitamos son programas informáticos que sean capaces de llevar a cabo estas tareas de forma automática. La mayor parte de estos programas utilizan los términos o palabras que componen los documentos para representar sus contenidos; y, también, las palabras que utiliza el usuario para expresar su necesidad informativa, esto es, cuando compone una consulta en lenguaje natural. Sucede que, con frecuencia, una misma palabra toma distintas formas muy parecidas, pero no exactamente iguales; los ordenadores no son inteligentes como las personas, y tienen cierta dificultad en entender que *biblioteca y bibliotecas,* por ejemplo, son casi lo mismo; y que, probablemente, si uno busca información sobre *catalogación* y en un documento se habla de *catálogos o de catalogadores,* probablemente trate de aquello que estábamos buscando.

Además, muchos de esos programas que venimos comentando basan su eficacia en que son capaces de discriminar la importancia de unos términos frente a otros. En un documento concreto (o en una consulta de un usuario), los términos no representan por igual el contenido temático de dicho documento (o consulta); unas palabras son más representativas que otras. Los programas asignan más o menos importancia o representatividad a cada término mediante diversas técnicas, pero la mayor parte de estas aplican cálculos que utilizan las frecuencias de aparición de las palabras en diversos contextos. Naturalmente, esas frecuencias no resultan igual si consideramos *biblioteca y bibliotecas,* o *catalogación, catálogo, catalogador* palabras distintas.

Desde hace ya bastantes años se pensó que se podrían mejorar los resultados de estos programas de recuperación si se conseguía agrupar bajo una única forma esas palabras que, con ligeras diferencias, hacían referencia a un mismo concepto. Los investigadores llamaron a esta operación *stemming*, en alusión a esa raíz común a las palabras que se agrupan. No hay una buena traducción al español, aunque muchos hemos optado por la expresión *lematización*; esta opción, por cierto, no tiene una aceptación unánime, y algunos de nuestros colegas se enfadan cuando la encuentran en artículos, ponencias y demás (hay quien piensa que el problema está en que algunos de nuestros colegas son demasiado propensos al enfado).

Sin embargo, se llame como se llame, no es una operación fácil para un programa informático. Se han ensayado numerosas técnicas, la mayor parte dependientes o válidas para un idioma concreto (el inglés en la mayoría de los casos, naturalmente). Para el caso del español ha habido pocos intentos, y de eso trata este libro: de la lematización en español.

Si comparamos el español (y, en general, las lenguas procedentes del latín) con otras, veremos que nos enfrentamos a una riqueza y complejidad morfológicas importantes. El ejemplo más evidente es el de los verbos irregulares; estos (y sus conjugaciones) son la pesadilla de cualquier estudiante de español como lengua extranjera, y en cualquier diccionario encontraremos varios millares de ellos, cada uno de los cuales puede dar lugar a una buena cantidad de flexiones (modos, tiempos, personas y número).

Del mismo modo, nos encontramos con la gran variedad de palabras derivadas que se forman añadiendo determinados sufijos a una raíz dada. El problema es que, además del número de sufijos que tenemos en la lengua española, estos se pegan a las raíces de las más variadas formas, modificando incluso dichas raíces hasta dejarlas, a veces, difícilmente reconocibles. De manera que no es posible abordar la lematización en español sin codificar el complejo y abundante conocimiento lingüístico que explica tanto las variaciones de las palabras a través de las flexiones morfológicas como a través de la derivación. Y esto es, precisamente, una parte importante de lo que Raquel Gómez nos presenta en este trabajo.

Pero es que, además, la presunción de que la lematización mejora los resultados de la recuperación no es más que eso: una mera presunción. Presunción, por cierto, que, aunque fundada, ha sido puesta en discusión en numerosas ocasiones. Nuevamente estamos ante una cuestión dependiente del idioma, de manera que lo que produce determinados resultados recuperando documentos en inglés no tiene por qué funcionar de la misma manera recuperando documentos en español. Así que no se trata solamente de lematizar, sino de aplicar la lematización a la recuperación de información. Esta es otra parte importante del trabajo de Raquel Gómez: la investigación experimental, a fin de evaluar el impacto o incidencia de la lematización en los resultados de la recuperación en español.

Y termino refiriéndome a las otras partes de este trabajo que, sin ser tan novedosas, resultan también inestimables. Estoy hablando de las partes previas o introductorias en las que, en pocas páginas, se nos presenta de una forma realmente