TÉCNICAS LINGÜÍSTICAS APLICADAS A LA BÚSQUEDA MULTILINGÜE: AMBIGÜEDAD, VARIACIÓN TERMINOLÓGICA Y MULTILINGÜISMO

Anselmo Peñas Padilla
Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 2004

Un profesor quiere acceder a *recursos* sobre *educación especial* [...] Podría emitir la siguiente consulta: *recurso educación especial*. El primer documento que recupera uno de los mejores buscadores en Internet es una orden ministerial que contiene: «... recurso contencioso/administrativo ... educación especial...».

Con este ejemplo comienza el autor la explicación de la ambigüedad léxica, los problemas que ésta plantea en la Recuperación de Información y distintas formas de abordarla. El tema central del libro son, precisamente, las barreras que el lenguaje impone en la Recuperación de la Información: polisemia (y sinonimia), variaciones morfosintácticas, semánticas o translingües de los términos o palabras que sirven tanto para definir el contenido temático de los documentos, como para formalizar las necesidades informativas de los usuarios.

Conviene precisar que el punto de partida es la consideración del proceso de Recuperación como un proceso totalmente automatizado: tanto en la indización o representación del contenido de los documentos, como en la expresión de las necesidades informativas (consultas) de los usuarios, como, desde luego, la estimación del parecido, cercanía o similitud entre consulta y documentos. El escenario, pues, es un sistema de recuperación de tipo *best-match*, en el que los documentos se indizan automáticamente y las consultas se formulan en lenguaje natural; el resultado de resolver una consulta, por otra parte, es una lista de documentos ordenada en función de la mayor o menor adecuación del documento a la consulta. En consecuencia, nada de indización manual, lenguajes controlados, y demás. La concepción eminentemente multidisciplinar de la Recuperación de Información es palpable a lo largo de todo el libro.

Desde este punto de vista, algunas de las partes consideradas como introductorias tienen un valor inestimable. Constituyen una explicación relativamente breve, pero de gran claridad, tanto de los problemas que debe afrontar la Recuperación de la Información, como de las técnicas más empleadas para resolver tales problemas. La brevedad no disminuye el rigor en la explicación, pero todo ello se hace de una forma asequible y fácil de seguir. En esta línea cabe destacar la exposición sobre los métodos de experimentación (y evaluación de los resultados de los experimentos) en Recuperación de la Información, la aplicación de diversas técnicas de Procesamiento de Lenguaje Natural, y, particularmente, la manera de abordar la Recuperación Multilingüe; es decir: aquellas situaciones en las que consultas y documentos utilizan diferentes lenguas, en ocasiones con combinaciones especialmente complejas; como, por ejemplo, documentos en diferentes lenguas, incluso con partes en distintos idiomas dentro de un mismo documento.

A partir de ahí, el autor describe diversos experimentos llevados a cabo para estimar la utilidad de aplicar diferentes técnicas lingüísticas al proceso de Recuperación (Automática, no olvidemos). Especialmente interesantes son los que utilizan *WordNet*

o, mejor aún, *EuroWordNet*. Y esto no sólo por los experimentos en sí, sino también por la información adicional que se proporciona sobre ambos repertorios terminológicos (si es que se les puede llamar así); los cuales, por otra parte, y sobre todo *EuroWordNet,* son bien conocidos por el autor.

La conclusión es que el uso de tales técnicas lingüísticas en un contexto de Recuperación Automatizada estándar es poco útil para mejorar los resultados de dicha Recuperación. A partir de aquí, el autor se plantea cuál deba ser el papel de dichas técnicas lingüísticas en el proceso de Recuperación. Esta reflexión le lleva a considerar otro escenario diferente en el cual se introduce la interactividad con el usuario; interactividad a la hora de ayudar a éste a expresar su necesidad de información, e interactividad también a la hora de navegar por los documentos obtenidos en la recuperación, a fin de obtener finalmente los más adecuados.

Es en este contexto de interacción donde las técnicas lingüísticas puede resultar útiles. Para corroborar esta hipótesis, el autor construye un sistema de Recuperación de Información que utiliza técnicas de Procesamiento de Lenguaje Natural para interactuar con el usuario. Este sistema, que se apoya en el motor de búsqueda *ITEM*, desarrollado por el *Grupo de Procesamiento de Lenguaje Natural de la UNED*, recibe el nombre de *Website Term Browser* (*WTB*). Diversos experimentos son llevados a cabo con él, a fin de demostrar la eficacia de esas técnicas de NLP en la interacción con el usuario.

El libro está editado por la Sociedad Española para el Procesamiento del Lenguaje Natural, que anualmente premia los trabajos originales de investigación en este campo; el trabajo que comentamos es uno de los premiados y, consecuentemente, editado por la SEPLN. El autor, por otra parte, es miembro del Grupo de PLN de la UNED, de dilatada experiencia en el campo del Procesamiento del Lenguaje Natural y de la Recuperación de Información; es especialmente conocida su participación en el proyecto *EuroWordNet*, así como su actividad de soporte y organización del *CLEF* (*Cross Lingual European Forum*), probablemente el foro o simposio europeo más importante sobre Recuperación de Información.

Se trata, obviamente, de un trabajo de investigación, que podría pensarse, en principio, no apto para no iniciados. Sin embargo, además de su interés para investigadores en el campo de la Recuperación de Información, la claridad en la exposición hace que, al menos la primera parte, sea también recomendable para quien simplemente quiera ponerse al día en lo que es la Recuperación de Información actual, de nuestros días; y lo es, para hacerlo con un texto en español, sin recurrir a intermediarios, sino utilizando como fuente a quienes directamente trabajan en ese campo.

Cargos G. Figuerola
Universidad de Salamanca, Grupo REINA