Diccionarios on line

F.de Moya,J.M.Muñoz,P.Hipola
Universidad de Granada

Son conocidas las prestaciones de los diccionarios "on line" en tanto que herramientas para el análisis automático de documentación textual, la elaboración de productos de indización, para la traducción asistida por ordenador (TAO) y para la preparación de léxicos especializados. Como útiles del análisis automático, los diccionarios pueden constituir, en primer lugar, instrumentos para automatizar los procesos de reducción de las diferentes formas lingüísticas flexionadas a sus términos lexicales canónicos, para que, a continuación, gracias a los pertinentes análisis estadísticos de frecuencias de vocabulario, se llegue a la determinación de la especificidad del contenido de los documentos, a través de técnicas usuales en "word indexing".

Por otra parte la posibilidad de que los diccionarios on line establezcan relaciones entre los significantes de dos

lenguas hace que sean ayudas al menos tan eficaces, pero sin duda más manejables y volubles, como los diccionarios convencionales. Y esto se manifiesta en mayor medida en el ámbito de los vocabularios de términos especializados, puesto que uno de los problemas fundamentales de la puesta al dia de los léxicos de materias especificas es la rápida evolución en nuestro tiempo de las ciencias experimentales y de la tecnologia.

En unión a todo esto, los sistemas de edición electrónica se encuentran hoy dia tan extendidos que es lógico pensar que la tarea de introducir grandes volúmenes de documentación textual en los sistemas automáticos, para mantener al dia y utilizar los distintos tipos de diccionarios on line, ya no es algo tan problemático como lo pudiera ser en el pasado.

Dentro de este mismo orden de problemas, se impone de modo urgente la revisión de los diccionarios de frecuencias que para la lengua standard existen, con el fin de poner a disposición de los especialistas un punto de referencia válido en este sentido.

Junto a todo este conjunto de factores, otra serie de circunstancias, de índole diversa, hacen que sea oportuno preguntarse cuáles pueden ser hoy las funciones, viabilidad, estructura, posibilidades y limitaciones de estos diccionarios en el contexto de los problemas de la lexicografia actual.

En primer lugar, hay que considerar que el universo de la informática en muy pocos años ha variado por completo la estructura de sus planteamientos. Por una parte, ha aumentado vertiginosamente la velocidad de proceso en los equipos clasificados dentro de la gama de los "ordenadores menores". Los

tiempos de espera para la respuesta en modo interactivo se han reducido de tal manera que ahora es oportuno plantearse que sea una de estas máquinas quien se ocupe de llevar a cabo la puesta en funcionamiento de software para el tratamiento de material lingüistico de propósitos ambiciosos. Por otra parte, la optimización de la circuitería y de los sistemas operativos se ha visto acompañada además por el crecimiento de la capacidad de los dispositivos de almacenamiento masivo, de forma que ya se puede encargar a un micro la gestión de una base de datos con un volumen de información de varios cientos de megabytes.

A todo ello hay que añadir la aparición de nuevos dispositivos en la periferia de estos equipos, tales como los lectores ópticos, diferentes aplicaciones de la tecnología láser; la comercialización de micros multipuesto, etc.; factores todos ellos que hacen posible contar en nuestros dias con el hardware y las herramientas de software de base necesarios para que la gestión de un diccionario "on line" de amplias prestaciones pueda ser realizada en un "desktop computer".

En el caso de que para estos propósitos se efectúe la opción por uno de los denominados "ordenadores personales", se pueden encontrar también disponibles aplicaciones de considerable potencia adecuadas al proyecto. Nos queremos referir a los paquetes de gestión documental. Este tipo de sistemas ha experimentado una progresiva mejora de prestaciones. Los primeros programas que se desarrollaron, inspirados fundamentalmente en las intuiciones y metodología de Hans Peter Luhn sobre el análisis léxico estadístico automático, fueron pronto

perfeccionados por sucesivas implementaciones de módulos de programación para tareas específicas de tratamiento lingüístico. Así, los cómputos de frecuencias léxicas absolutas se complementaron con diversas técnicas:

-ponderación de las palabras según su posición en el texto,

-criterios de frecuencia léxica relativa y grados de probabilidad de aparición de las formas,

-utilización de listas de palabras no significativas (antidiccionarios),

-selección de vocablos correspondientes a determinadas categorías gramaticales,

-empleo de tablas de sinónimos y de thesaurus,

-descomposición morfológica en raíz-desinencias,

-rastreo de coocurrencias (por proximidad),

-aplicación en general de los procedimientos usuales en lingüística estadística: ley de Zipf, etc.

Las últimas generaciones de paquetes de gestión documental han procurado desarrollar algoritmos de interpretación sintáctica y semántica, de acuerdo con modelos lingüísticos especialmente preconcebidos, entre los que cabe recordar aquellos que componen los llamados analizadores morfográficos, encargados del reconocimiento de los componentes formales de la cadena textual. Tras realizar el análisis, el gestor en cuestión se puede ocupar de confeccionar y mantener al día los ficheros invertidos, compuestos por las entradas léxicas -formas lexicales canónicas y ciertos sintagmas completos- y los punteros necesarios, ofreciendo la posibilidad de establecer distintas indexaciones y

reindexaciones de los items de acuerdo con los diferentes criterios posibles: orden alfabético, relaciones jerárquicas o de red.

Junto con estas prestaciones hay que considerar la facilidad para operar, en modo interactivo o a través del diseño de procedimientos, con operadores booleanos. Y, en definitiva, el poder disponer del lenguaje lógico-formal imprescindible para la gestión inteligente de indices, que cuente con las prestaciones de los modernos lenguajes de programación estructurada y recursiva.

Para llevar a cabo la elección de un programa que se encargue de la gestión de un diccionario on line seria necesario escoger uno entre los que dispongan de los más versátiles sistemas de recuperación de información y que al mismo tiempo sea lo más "amigable" posible con el usuario en lo que se refiere al diseño y mantenimiento de la base de datos, puesto que tendria que poder ser utilizado por personas no especializadas en el ámbito informático. Asi, por ejemplo, sería conveniente que ofreciera facilidades como la edición por pantallas completas, displays rápidamente interpretables, etc. Además, es importante que el software de base que soporte el paquete tenga capacidad para relacionarse con otras aplicaciones, ya que, por su carácter instrumental, el diccionario que se cree ha de poder ponerse en funcionamiento desde diferentes entornos de software básico y de aplicaciones tales como tratamientos de textos o diferentes programas de usuario. Es inconcebible que la elaboración de un diccionario on line constituya una aplicación cerrada en si

misma.

Enunciadas estas consideraciones de carácter general, esbozaremos unas notas sobre algunas de las particularidades que, según nuestro parecer y de acuerdo con los resultados que arrojan las experiencias que venimos realizando desde hace algún tiempo, debería reunir un diccionario on line que satisfaga las funciones que se han enumerado en los primeros párrafos de estas páginas. En lo que respecta a la estructura interna del diccionario, la base de datos tendría que estar compuesta por una secuencia de registros, cada uno de los cuales constara de: a) la entrada, que sería el campo clave principal; b) datos de tipo estadístico -frecuencias relativas-; y c) datos de tipo gramatical y semántico: categoría, marca de "palabra vacía" sí/no, punteros que señalen a la lexical canónica correspondiente -si es el caso-, indicadores de "sentido".

La primera cuestión que se plantea al llegar a este punto es si en el conjunto de las entradas, junto a las lexicales canónicas se deben incluir formas flexionadas. Para que pueda llevar a cabo la reducción de todas las posibles formas a sus términos lexicales canónicos, cabría pensar que bastaría con que dispusiera de los algoritmos propios del analizador morfográfico. En el caso de trabajar con lengua castellana, nuestra opinión es que, además de utilizar estas rutinas, lo más efectivo es que en la base de datos se incluyan todas las flexiones posibles que puedan crearse a partir de cada una de las lexías y de cada lexema. Hacer uso de este procedimiento podría parecer hace unos años algo desproporcionado, pero hoy, con el abaratamiento de la

memoria de masa y la rapidez de los sistemas de gestión de índices, el planteamiento puede ser distinto. Y -esto es lo esencial- para algunas situaciones en el análisis, parece que éste es el único método posible.

Las informaciones de tipo estadístico son especialmente útiles para mantener actualizados de forma permanente los vocabularios de términos específicos de las diferentes áreas temáticas. Interesa que determinados elementos léxicos -formas, términos lexicales canónicos, lexias, campos semánticos- lleven asignados códigos que informen sobre la utilización peculiar de dicho elemento en textos de una materia especializada, con su frecuencia relativa en ese ámbito. El proceso de asignación de códigos puede efectuarse de manera automática por el sistema, que comparará las frecuencias relativas de utilización de los elementos en cuestión dentro de la documentación especializada con las correspondientes de la lengua standard.

En lo que se refiere a la caracterización semántica sería interesante organizar un completo subsistema -una especie de diccionario paralelo- de codificación de áreas léxicas en torno a lexias, y, en un espectro más amplio, en torno a campos semánticos. La organización del sistema de punteros de acuerdo con criterios de este tipo aparece como imprescindible si se ha de emplear el diccionario para determinar especificidades en el contenido significativo de los documentos dentro de un proyecto global de análisis e indización documentales automáticos.

Por último, cabría señalar que las tareas de confección de diccionarios on line no son contrarias, en nuestra opinión, al

trabajo que se realiza en IA para la creación de intérpretes de lenguaje natural. Dentro de la base de conocimientos que maneje cualquier sistema de este tipo ocupará una parte principal el diccionario. Mientras se avanza en el desarrollo de estos sistemas, se debe seguir ampliando la información de los diccionarios y sobre todo ponerlos a disposición de todos los usuarios potenciales.