PROCESAMIENTO DE TEXTOS EN EL CENTRO DE INVESTIGACIÓN UAM-IBM

Luis de Sopeña

Centro de Investigación UAM-IBM
Universidad Autónoma de Madrid

INTRODUCCIÓN

El Centro de Investigación UAM-IBM ha venido desarrollando a lo largo de los últimos años diversos proyectos relacionados con el Procesamiento Automático de Textos; se ha puesto especial énfasis en los problemas específicos que presenta el idioma castellano y sus posibles soluciones.

El interés y la necesidad de tales investigaciones son claros. No olvidemos que, todavía hoy, los dispositivos periféricos como teclados, pantallas e impresoras no disponen en muchos casos de caracteres extraños (por inexistentes) a la lengua del país en el que fueron diseñados. Esto implica, en nuestro caso, que el tener que escribir un texto en el que los símbolos propios de nuestra lengua sean correctamente aceptados y reproducidos por el ordenador se convierta en un problema considerable. Ello ocurre en castellano con las vocales con acento agudo, la ü, la ñ, y los signos de abrir exclamación e interrogación, así como en catalán con sus caracteres específicos: los mismos que en castellano y además, las vocales a, e y o con acento grave, la ï, la ç y el punto "alto" de la l geminada. (Sólo en algunos ordenadores personales o en máquinas específicas para

proceso de textos se ha tenido en cuenta este problema, pero nunca hasta ahora en los grandes ordenadores de propósito general).

Este es solamente un ejemplo trivial de lo mucho que queda aún por hacer en el tema del procesamiento de textos para disponer de un sistema completo y efectivo para las lenguas de nuestro país.

A continuación se describen brevemente dos de los proyectos incluidos en este área en los que el Centro Científico de IBM está trabajando en la actualidad: por un lado, la elaboración de un diccionario para corrección y ayuda a la escritura de documentos; por otro, un editor-formateador de textos en línea que permite ver en pantalla, "sobre la marcha", la forma final del documento que se está elaborando.

DICCIONARIO DEL CASTELLANO

Entendemos aquí por diccionario una colección de palabras (en principio, sin ninguna definición asociada), mejor cuanto más extensa, que indica la ortografía correcta de cada término y que adicionalmente puede llevar otras informaciones, tales como parte de la oración, género, número, tiempo, persona, etc.

La elaboración de un diccionario del español en el ordenador presenta un interés múltiple, ya que, aparte de su aplicación inmediata a sistemas como el que aquí se presenta, constituye el presupuesto básico para numerosas otras aplicaciones en que intervenga el lenguaje. El diccionario que hemos construido se aplica ya a la corrección ortográfica de cualquier documento o texto escrito en castellano: una carta, un artículo, un informe, un libro, etc.

Hoy en día existen herramientas eficientes y cómodas de utilizar para:

1. introducir un texto en el ordenador (programas editores, con facilidades de inserción, borrado, desplazamiento, copia, etc. y de los que se han desarrollado ya versiones de un considerable grado de sofisticación).

2. dar formato final a dicho texto para imprimirlo adecuadamente, por ejemplo, a base de intercalar marcas que describan cada parte del documento (cabeceras, títulos, notas a pie de página, listas), o que indiquen las características del formato que se desea; será después el programa quien, de acuerdo con unas normas predefinidas (e incluso modificables/ampliables por el usuario a fin de que pueda definir su propio "estilo"), pagine y dé formato al texto: cambio en los tipos de letra, alineación a la derecha, numeración de páginas, títulos y subtítulos, inserción de notas, composición de índices, localización de referencias a figuras, capítulos, etc. Ejemplos: DCF [1], JANUS [2].

(Como ejemplo, el presente artículo ha sido formateado con un programa de este tipo, concretamente DCF, e impreso a continuación en una impresora de láser, que dispone de todo el repertorio de caracteres españoles).

Es la etapa intermedia a las dos anteriores la que en la actualidad dispone de menor número de herramientas, y es a ella a la que se aplica nuestro trabajo. Las referencias [3], [4] y [5] presentan algunos de los sistemas más interesantes diseñados para estos propósitos en los últimos tiempos.

Objetivos

La aplicación principal del diccionario es la corrección morfológica de textos, considerando cada palabra de forma independiente. Otros tipos posibles de

corrección (sintáctica, semántica, estilística) no han sido considerados por el momento, pero podrían serlo más adelante.

Es decir, el prototipo construido trata, fundamentalmente, de aislar errores tanto ortográficos como de pulsación en los textos que se le presentan a examen, en definitiva, palabras escritas incorrectamente.

Utilización

En cada documento que le es sometido, el programa detecta y marca (iluminándolas con otra intensidad - más brillante, en negativo, parpadeante - en una pantalla monocroma, o empleando un color diferente, en el caso de un monitor de color) aquellas palabras que no encuentra en el diccionario: el motivo puede ser tanto un error ortográfico como una trasliteración al teclear, o bien que la palabra, siendo correcta, no está en el diccionario.

En este último caso el programa permite al usuario fabricar un apéndice en el que introducirá el término; este apéndice será en lo sucesivo consultado al igual que el diccionario principal. De este modo, junto con el diccionario base proporcionado con el programa, el usuario puede fabricar sus propios glosarios de términos técnicos, nombres propios, siglas que emplee habitualmente, etc.

En los otros casos, si el usuario no está seguro o desconoce la ortografía correcta de la palabra, puede solicitar ayuda del programa, quien le proporcionará uno (o varios - hasta un máximo de 6) términos que sí están en el diccionario y son muy parecidos al erróneo. El "parecido" entre dos palabras es determinado por un algoritmo y esencialmente depende, de forma inversamente proporcional, del número de alteraciones que es preciso operar sobre la palabra del diccionario para obtener la palabra equivocada.

La Figura 1, al final del artículo, muestra la imagen de una pantalla en que aparece un texto procesado por el programa. Las palabras erróneas son resaltadas y, al solicitar candidatos en sustitución de un error, se abre una "ventana" con las palabras correctas. Un menú en la parte inferior de la pantalla informa en cada momento de las distintas alternativas programadas en las teclas de función.

Estructura

A la hora de realizar una verificación ortográfica puede optarse por una de dos soluciones: un programa de análisis morfológico basado en un diccionario de raíces y excepciones, o bien un diccionario que incluya todos las posibles formas flexionadas de una palabra. En nuestro caso, la solución adoptada ha sido la segunda, utilizando una serie de mecanismos que proporcionan una compactación máxima, lo que garantiza una ocupación mínima de espacio de memoria.

Para ello ha sido preciso realizar un análisis morfológico "manual" previo, al objeto de clasificar las palabras de acuerdo con los diferentes grupos de terminaciones, que también ha sido necesario determinar. El problema es especialmente agudo en los verbos, donde ha habido que reclasificar las conjugaciones e irregularidades presentes en la literatura, e introducir otros, no considerados habitualmente como tales, pero necesarios en virtud del tratamiento informático. Asimismo, ha sido preciso estudiar los problemas planteados por los pronombres enclíticos para su introducción en el diccionario. Una descripción detallada de las soluciones adoptadas para los verbos, los tipos de irregularidades considerados y una relación clasificada de los mismos puede encontrarse en [6].

Después, a base del fichero de colecciones de terminaciones y del de entradas principales (que contendrá apuntadores al anterior), se construye el diccionario

rio siguiendo un proceso de compactación. Este será el utilizado por el programa de verificación. Del grado de compactación alcanzado pueden dar una idea las cifras siguientes: actualmente el diccionario consta de unas 37.000 entradas, que flexionadas dan lugar a 426.000 palabras distintas; pues bien, el tamaño total ocupado es de 305 kbytes.

En cuanto a las palabras incluidas, se ha procurado cubrir hasta el máximo el lenguaje escrito usual, quedando por tanto excluidos términos técnicos, coloquiales o inusuales. Se encuentran por un lado los grupos "acotados" de palabras: artículos, preposiciones, conjunciones, adverbios, numerales, pronombres, adjetivos posesivos y demostrativos, nombres de días y meses, nombres propios de persona, apellidos, nombres geográficos, etc. Por otro lado, los conjuntos "abiertos" de sustantivos, adjetivos calificativos y verbos.

La información adicional que el diccionario incluye para cada entrada, en estos momentos, consta de los puntos de corte silábico y de la parte de la oración, es decir, la función morfológica.

Próximas tareas

En la actualidad se está trabajando en la ampliación del diccionario en dos direcciones, con la introducción de sinónimos y de las que se han llamado palabras confundibles.

Un diccionario de sinónimos integrado en el anterior permitiría proponer al usuario que los consultase sinónimos para una palabra dada, con el objeto, por ejemplo, de perfeccionar, matizar o mejorar estilísticamente la redacción de un texto.

Las palabras confundibles son un intento de afinar la verificación ortográfica. Llamamos confundibles a los grupos de palabras (usualmente parejas o tripletes) que están en el diccionario pero son homónimas o parónimas (ejemplos: vaca/baca, arte/harte, uve/hube, callado/cayado, ora/hora, aya/haya, etc.) y los pares de palabras con y sin acento, muy numerosos en nuestro idioma (que/qué, si/sí, el/él, se/sé, de/dé, etc. y especialmente tiempos de verbos: amo/amó, amara/amará, etc.); en este último caso, el intento fundamental sería avisar o corregir el posible olvido del acento.

En algunos casos, el propio programa puede dilucidar si la palabra "confundible" está bien escrita, desambiguando a base de estudiar el contexto en que aparece. En los demás casos (error o imposibilidad de decidir), las palabras detectadas como "confundibles" se resaltan para que el propio usuario decida (proponiéndole las distintas alternativas, si así lo desea), candidatos confundibles con el término que ha utilizado.

Otra posibilidad que también contemplamos para el futuro es la realización de un analizador sintáctico de las frases del texto, basado en el diccionario, y que tendría nuevas y diferentes implicaciones en cuanto a la verificación del documento.

EDITOR-FORMATEADOR

Se trata de un programa que combina las facilidades de un editor con las de un formateador "en línea" (junto con, adicionalmente, verificación ortográfica basada en el mismo diccionario anterior). El objetivo fundamental es que la forma del documento tal como se va viendo en pantalla a medida que se va intro-

duciendo, sea lo más parecida posible (si no exactamente igual) a como quedará finalmente tras de ser impreso. El soporte físico necesario para ello deberá disponer de un terminal capaz de visualizar distintas "fonts" y caracteres programables, así como de una impresora de láser.

Frente a un formateador como el anteriormente descrito, que trabajaría en modo "batch", una vez completado el fichero de "datos" (texto) e "instrucciones" (órdenes de formato y descriptores del texto), el que aquí se propone funcionaría de modo totalmente interactivo. A medida que se van introduciendo las distintas partes del texto o documento, se va indicando su descripción, es decir, si se trata de un título, de un nuevo párrafo, si hay un cambio en el tipo de letra, etc. Para ello se utilizan continuamente las teclas de función. La pantalla tiene siempre un espacio reservado para "ayuda" al usuario sobre el significado de las teclas especiales; ello hace muy sencilla su utilización y muy rápido el habituarse a esta forma de trabajar.

Otra característica interesante del programa es que dispone de rutinas especiales para trazar cajas y tablas y para componer fórmulas matemáticas de complejidad arbitraria.

Una vez que el documento ha sido introducido en el ordenador, paginado y corregido (cosas que, como hemos dicho, ocurren simultáneamente), puede enviarse a la impresora con la seguridad de que el resultado impreso será una "fotografía" de lo que se ha ido produciendo en la pantalla (con las únicas diferencias que pueda haber entre los tipos de font, espacio compensado, número de pels, ... debidos a las resoluciones respectivas y a las limitaciones propias de cada dispositivo).

REFERENCIAS

- [1] Document Composition Facility - Script/VS Text Programmer's Guide, IBM Program Number 5748-XX9.

- [2] JANUS User's Guide, IBM San Jose Research Laboratory, 1983.

- [3] SIGPLAN Notices, vol. 16, no. 6, June 1981.

- [4] IBM Systems Journal, vol. 21 no. 3, 1982.

- [5] J. André: Bibliographie analytique sur les "manipulations de textes", Technique et Science Informatiques, vol. 1 no. 5, 1982.

- [6] R. Casajuana, C. Rodríguez: Verificación ortográfica en castellano. La realización de un diccionario en ordenador, próxima publicación, disponible internamente.

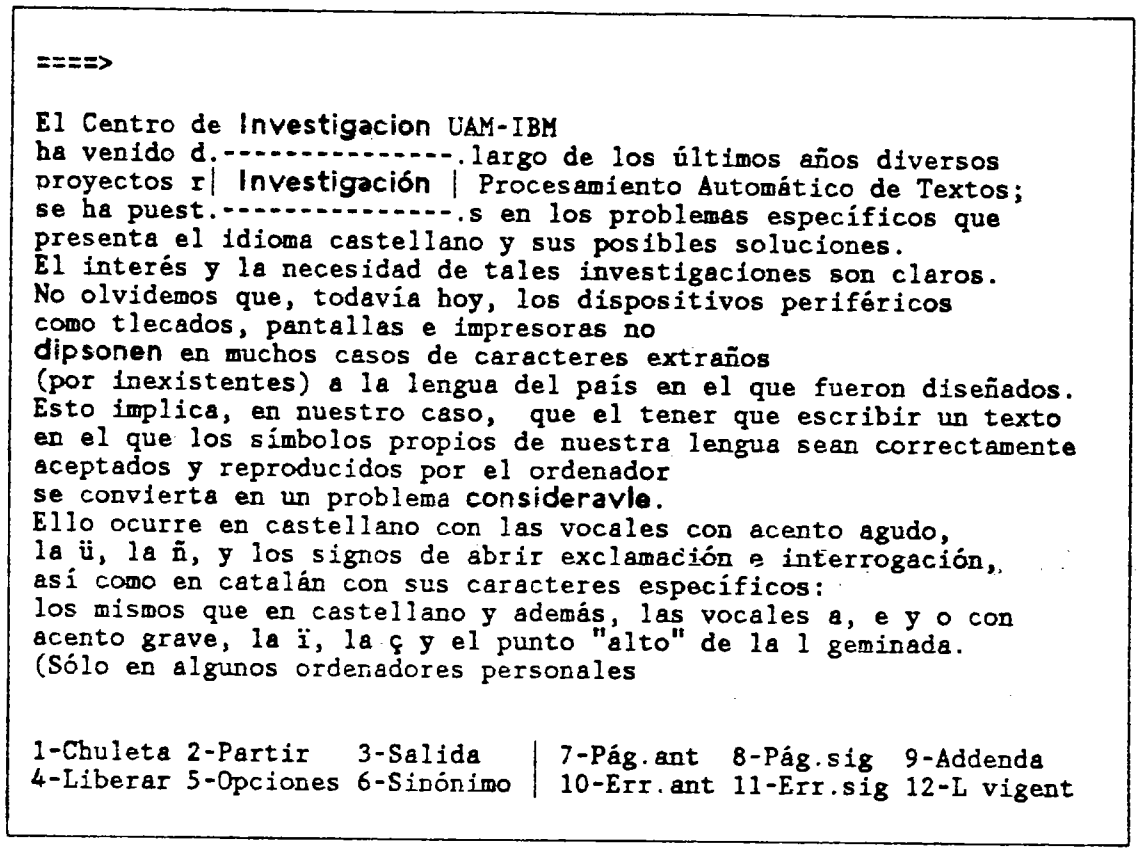

Figura 1. Ejemplo de uso del diccionario