La recuperación automática de la información

Avances en el tratamiento de textos en español

Antonio Moreno Sandoval

La colaboración entre la lingüística y la informática está permitiendo el tratamiento y el análisis de grandes colecciones de datos textuales. El proyecto PROTEUS y sus aplicaciones muestran los avances realizados con textos en español.

1. ¿QUÉ ES EL PROCESAMIENTO INFORMÁTICO DE LAS LENGUAS NATURALES?

Desde hace más de tres décadas se está trabajando en lo que se conoce como Lingüística Computacional o también Procesamiento del Lenguaje Natural: una disciplina aplicada que reúne los contenidos de dos ciencias, la Lingüística y la Informática. La idea de que los ordenadores puedan *entender* realmente el lenguaje humano (y no simplemente instrucciones que deben ser escritas de una manera rígida y concreta) ha estado en la mente de todos desde que los ordenadores comenzaron a intervenir en nuestras vidas. Incluso la literatura y el cine nos han sugerido la posibilidad de poder conversar con máquinas inteligentes (recordemos por ejemplo el ordenador central de 2001, *una odisea del espacio , o* el robot parlanchín de *La guerra de las galaxias).* No cabe duda de que no podremos disponer de semejantes *colaboradores* en un futuro inmediato, pero en cambio muchas actividades podrán realizarse -de hecho se pueden realizar ya- sin el esfuerzo y la dedicación de un ser humano. Nos referimos concretamente a aquellas actividades donde sea necesario el tratamiento de la información codificada en una lengua natural como el español, el inglés o el chino. Por el contrario, se entiende por *lenguas artificiales o formales* aquellas que han sido creadas por el hombre para formalizar el conocimiento y poder llevar a cabo operaciones con él, por ejemplo, el lenguaje matemático o los lenguajes de programación. La diferencia fundamental con las lenguas naturales es que los lenguajes artificiales carecen de ambigüedad y su sintaxis es mucho más rígida.

Debido precisamente a la flexibilidad y riqueza de las lenguas naturales su tratamiento por ordenador se hace muy complejo y es necesario, por tanto, restringir el campo de aplicación a dominios lingüísticos concretos en busca de patrones sintácticos y semánticos más rígidos que permitan su interpretación de una manera inequívoca. Por ejemplo, en el lenguaje oral se permiten muchas más libertades expresivas que en el escrito: uno puede dejar oraciones incompletas o utilizar palabras aproximadas con la seguridad de que el contexto ayudará a su(s) oyentes) a entender las ambigüedades e imprecisiones. En un texto escrito -cuando menos- las oraciones tienen que ser gramaticales, y si queremos que nuestros lectores nos entiendan debemos ajustarnos lo más posible a una única interpretación. Consecuentemente, los textos ambiguos, con múltiples lecturas, no son aptos para ser tratados por ordenador. Típicamente son los textos literarios donde el autor juega con el lenguaje de una manera artística. Por el contrario, los textos técnicos y científicos son apropiados para ser interpretados automáticamente: el autor utiliza un lenguaje sin ambigüedades, donde no pretende decir nada más que lo que realmente dice. Normalmente, el número de construcciones sintácticas (sintagmas, oraciones, párrafos) y de palabras no es excesivo, y el vocabulario técnico no es ambiguo: cuando decimos que "el limitados salta cuando se sobrepasa un determinado límite de cargó", nos referimos a una de las acepciones del verbo saltar (concretamente, la que no es sinónimo de brincar, dar saltos) y a una acepción particular de carga (la cantidad de electricidad que está soportando el circuito, y que no es equivalente a peso).

Tenemos, por tanto, que un texto ideal para ser tratado por ordenador podría ser un manual de instalaciones eléctricas, por ejemplo. Su vocabulario es inequívoco y sus construcciones sintácticas no son complejas. En resumen, se puede conocer la información que contiene sin mucho margen de error. Dada esta característica, podríamos desarrollar distintos sistemas informáticos que utilizaran esta información. Básicamente, podríamos tener dos aplicaciones: una sería traducir dicho manual a otra lengua -lo que se conoce por traducción automática (1)-; la otra, extraer la información y exponerla en un formato que sea más rápido de leer y consultar (por ejemplo, en forma de registro de base de datos o en forma de plantilla). A esto último se lo denomina extracción o recuperación de información. Si desarrolláramos estos sistemas pensando en traducir o interpretar unos cuantos manuales, obviamente los resultados no compensarían el esfuerzo y la inversión. Pero si trabajáramos con un número grande de textos (o información escrita en otros formatos) el ahorro de tiempo y dinero sin duda seria considerable y, por tanto, la inversión podría ser rentable.

Las administraciones, públicas y privadas, trabajan con enormes cantidades de textos y algunas cuentan con sistemas informáticos que les ayudan a manejarlos de una forma mucho más eficiente.

Por ejemplo, la CEE utiliza sistemas de traducción automática para traducir sus documentos a las nueve lenguas oficiales (aunque siempre se requiere la corrección a posteriori de los textos traducidos mecánicamente). El Gobierno americano está desarrollando sistemas para extraer información de textos periodísticos, de manera que la información clave se muestre en unas tablas. No son más que ejemplos de aplicaciones que se utilizan y que se utilizarán en el futuro cercano.

En la actualidad, la aplicación de las últimas innovaciones en el campo del procesamiento de lenguas naturales para el tratamiento automático de grandes colecciones documentales es uno de los objetivos prioritarios de los planes de I+D en el área de tecnología de la información, no solamente a nivel nacional sino especialmente a nivel internacional, donde los países más avanzados llevan investigando desde los años 60.

En resumen, la Lingüística Computacional es una ciencia aplicada (o ingeniería) que se encarga del desarrollo de sistemas informáticos que comprendan las lenguas naturales. Entre otras aplicaciones, hemos citado la traducción automática y la extracción de información pero también se incluyen los interfaces para consultar bases de datos utilizando una lengua natural, o los populares correctores ortográficos, gramaticales y de estilo.

2. EXTRACCIÓN DE INFORMACIÓN DE TEXTOS

Ya hemos hablado de que gran cantidad de información sólo está disponible en forma escrita: manuales, informes técnicos, documentos legales, noticias de periódicos, etc. Muchas veces necesitamos acceder a cierta información que está *escondida entre* montañas de documentos de una forma rápida y eficiente. Evidentemente, una manera es leerse cada uno de los documentos y comprobar por uno mismo si su contenido nos interesa, pero esto es sin duda costoso. Mucho más útil es tener almacenada una porción de la información total (es decir, la información más relevante) en una forma más estructurada -por ejemplo, en una base de datos convencional- de tal manera que nuestro acceso al contenido de cada documento sea notablemente más rápido.

El objetivo fundamental es tratar de emular la capacidad humana de interpretación de mensajes escritos mediante el uso de programas informáticos. Como cualquier otro tipo de automatización, estos sistemas computacionales liberarán a los especialistas humanos de muchas tareas repetitivas y que exigen, por otra parte, gran esfuerzo de concentración. Una ventaja adicional es que los ordenadores pueden funcionar sin descanso, consiguiendo resultados que sólo se lograrían con una fuerte inversión de personal y tiempo. La característica más sobresaliente de los sistemas de extracción de información es que permiten la cooperación, o, mejor dicho, la combinación de las habilidades más apropiadas de los humanos y de las máquinas: los analistas humanos son claramente superiores a los ordenadores en tareas complejas como la interpretación de información ambigua. En cambio, las máquinas pueden aventajar a los especialistas en tareas que requieren un alto grado de concentración y atención, como por ejemplo buscar en amplias cantidades de textos con baja densidad de información. En estos casos, es frecuente que pase desapercibida información relevante escondida entre montones de datos prescindibles. La tarea de estos sistemas será, por tanto, procesar previamente los textos para filtrar la información relevante de la irrelevante, dejando que los analistas humanos se concentren en las tareas complejas y altamente especializadas. La meta de algunos proyectos informáticos de los últimos años ha sido precisamente desarrollar sistemas de este tipo.

Concretamente, en el *New York University* (NYU) se está trabajando en este campo desde mediados de los años 70 y en la actualidad disponen de un sistema llamado PROTEUS *(PROtotype TExt Understanding System)* para analizar y extraer información de textos escritos en inglés. Dicho sistema tiene una cobertura bastante amplia en cuanto a construcciones sintácticas del inglés y su diccionario contiene alrededor de 35.000 entradas léxicas (equivalentes a las entradas de un diccionario impreso). Los autores del artículo han desarrollado un sistema similar para el español. En la actualidad cuenta con una cobertura sintáctica bastante similar a la del inglés, aunque con un diccionario mucho menos elaborado. El dominio temático de aplicación en ambos casos es interpretar textos periodísticos, aunque solamente los informativos y no los artículos de opinión.

A diferencia de otros sistemas de recuperación de información, nuestro sistema no selecciona documentos (o fragmentos de documentos) que pueden contener la información requerida, sino que *resume* el contenido de los documentos y lo muestra de una forma muy estructurada y accesible, a la que posteriormente se le puede aplicar un proceso de recuperación de información.

Otro aspecto importante es que el usuario de PROTEUS puede modelar el tipo de información que considera relevante. Aunque esto requiere que el dominio temático de los textos esté muy nítidamente acotado y que la estructura de la base de datos se determine antes del procesamiento de los textos. Esto implica que el sistema es reutilizable para diferentes dominios temáticos, siempre que los modelos interpretativos se adapten específicamente a los nuevos temas. Este tipo de sistemas se adapta

idealmente a textos y documentos de tipo técnico, como por ejemplo informes médicos, manuales de funcionamiento, reportajes científicos y de medio ambiente, textos jurídicos y administrativos (boletines oficiales, etc.) y textos periodísticos de carácter informativo.

En resumen, los sistemas de extracción de información facilitan el acceso y tratamiento de grandes colecciones de datos textuales, y mejoran la productividad en las tareas de información y análisis.

3. ESTADO ACTUAL DE PROTEUS

Los orígenes del proyecto PROTEUS datan del otoño de 1984. El Prof. R. Grishman, del departamento de Informática de la Universidad de Nueva York, desarrolló un analizados sintáctico que sirviera como base común para todas las aplicaciones que se crearan dentro del proyecto. Muchos aspectos del diseño del sistema reflejan la herencia del famoso y legendario *Linguistic String Project,* desarrollado (y todavía en uso) por este departamento desde mediados de los años 60 (Sager 1981). El sistema actual incluye un analizados léxico y otro semántico, además del sintáctico, y un generador de plantillas (o registros de bases de datos) especialmente diseñado para la aplicación en extracción de información. El proyecto PROTEUS cuenta con varias aplicaciones, entre ellas la consulta a bases de datos utilizando el inglés para comunicarse con el ordenador, pero sobre todo destaca por su participación en todas las conferencias que ha organizado el Gobierno americano sobre la extracción de información (conocidas como *Message Undestanding Conferences, MUG).* En las cuatro conferencias que se han convocado desde 1987, PROTEUS se ha situado siempre entre los cinco primeros grupos de investigación en Estados Unidos en esta área. El objetivo de estas conferencias, organizadas y subvencionadas por DARPA *(Defense Advanced Research Projects Agency),* persigue la evaluación de las distintas tecnologías existentes actualmente en el ámbito de la investigación avanzada en sistemas inteligentes.

El sistema PROTEUS fue desarrollado inicialmente para analizar textos en inglés. En los últimos años se ha extendido también al japonés y al español. La versión española ha sido desarrollada por los autores del artículo durante su estancia de 16 meses en la NYU. Varios artículos y conferencias recogen los resultados de la investigación, que se pueden resumir en los siguientes puntos:

1. El sistema PROTEUS ha demostrado su capacidad de trasladarse a otras lenguas, con resultados similares a los obtenidos en inglés.
2. El sistema PROTEUS en su estado actual de la versión inglesa consigue extraer entre el 40 y el 50 por ciento de la información relevante en textos sobre terrorismo en Hispanoamérica (dominio temático sobre los que se aplicó la evaluación de la última conferencia, MUC-4), con una precisión también en torno al 50 por ciento.

4. UN EJEMPLO CONCRETO

En este apartado presentaremos, de manera simplificada, el funcionamiento del sistema ante un caso real.

Cuando nos enfrentamos con un texto encontramos diferentes problemas que deben resolverse en sucesivas etapas. En primer lugar, tenemos un input que es simplemente una cadena de palabras en la que el ordenador no reconoce ninguna estructura. Por lo tanto, nuestra primera tarea es determinar la estructura de las oraciones, es decir, reconocer las relaciones que existen entre las palabras. Técnicamente, esto se conoce como *análisis sintáctico.* Por ejemplo, en "Para las puestas a tierra el instalador empleará principalmente electrodos artificiales", tenemos que reconocer que *empleará* es el verbo, *el instalador* es el sujeto, y electrodos artificiales es el objeto.

Una vez reconocida la escritura, hay que reconocer el significado de la oración, *o análisis semántico.* En esta fase del procesamiento se nos presenta un problema típico de las lenguas naturales: una misma idea puede ser expresada de varias formas, o mejor dicho, con distintas estructuras sintácticas u oraciones. Continuando con nuestro ejemplo, podemos decir también: "electrodos artificiales serán empleados por el instalador para las puestas a tierra" o "se emplearán electrodos artificiales para las puestas a tierra" (en este último caso, si queremos evitar decir quién fue el autor de la acción). En las tres oraciones existe la misma estructura semántica: una acción (que se corresponde semánticamente con el verbo), un agente (el instalador), un instrumento (electrodos artificiales) y un tema (las puestas a tierra). De la misma manera que podemos reducir las tres oraciones a una única estructura semántica, nos interesa que este tipo de expresiones se almacenen de una única forma en la base de datos. Es por ello que en nuestra plantilla informativa se representa la información de una manera más abstracta, donde los detalles superficiales (como el adverbio *principalmente)* son eliminados. Un ejemplo de plantilla sería:

Tipo de acción: Emplear
Autor de la acción: El instalador
Tema de la acción: Puestas a tierra
Instrumento: Electrodos artificiales

Si trabajáramos con oraciones aisladas habría que producir plantillas o registros de la base de datos para cada oración. Evidentemente con esto no conseguiríamos nuestro objetivo de *resumir* el contenido de un artículo. Por ello, lo que hacemos es, dependiendo del tipo de tema del artículo, utilizar plantillas generales con bastantes campos (o huecos para rellenar), elaboradas a partir del estudio semántico de posibilidades. Es decir, se escoge un tema concreto y bien delimitado, y se estudian los campos de registro (cada uno aportando una información relevante) que pueden aparecer. La experiencia ha demostrado que no se necesitan muchos campos (entre 15 y 20) para dar cuenta de la información importante sobre un tema particular. Por supuesto, muchos de los campos pueden quedar vacíos, pues no es habitual que en un artículo se cubran todos los aspectos del tema en cuestión.

5. PERSPECTIVAS

No es nuevo en Lingüística Computacional el hecho de que cualquier buen sistema de comprensión de lenguas naturales requiere una base rica de conocimiento del *mundo,* es decir, saber inferir interpretaciones correctas sobre ideas que no aparecen explícitamente en el texto. Siguiendo con nuestro ejemplo del limitados, nuestro conocimiento sobre instalaciones eléctricas nos permite interpretar correctamente palabras con más de un significado. De lo expuesto se puede deducir que cuanto más general queramos que sea nuestro sistema de extracción de información, más conocimiento del mundo debemos recoger, al tiempo que debemos contar con un mecanismo de razonamiento que sepa utilizar dicho conocimiento. Esto no quiere decir que el procesamiento automático de lenguas naturales sea una utopía: simplemente tenemos que limitar el conocimiento que se necesita para interpretar textos, restringiéndonos a dominios muy concretos.

Por otra parte, hay que aceptar cierto grado de error en los resultados de estos sistemas. Como en muchos casos el volumen de textos es demasiado grande para ser procesados directamente por seres humanos, parece preferible conseguir resultados parciales que no obtener nada. Además, los resultados obtenidos por analistas humanos distan mucho de ser excelentes. En la MUC-4 se realizó el siguiente experimento: dos analistas especialistas en el tema de la evaluación trabajaron sobre los mismos textos sobre los que lo hicieron los sistemas informáticos, con el fin de comparar la actuación humana y la computacional. Los analistas no consiguieron recuperar más del 75 por ciento de la información relevante de los textos, mientras que el mejor sistema se acercó al 60 por ciento (Sundheim 1992).

La falta de resultados espectaculares en este tipo de sistemas ha frustrado muchos intentos, pero es innegable que se van haciendo progresos cada año y, sobre todo, se puede hablar ya de programas que funcionan satisfactoriamente ayudando en distintas actividades humanas.

Notas

(1) A. MORENO SANDOVAL ha publicado en colaboración con F MARCOS MARÍN y F SÁNCHEZ LEÓN un extenso artículo sobre el proyecto EUROTRA de traducción automática en el número 16 (1988) de Telos, págs 90-99.

REFERENCIAS BIBLIOGRÁFICAS

El proyecto PROTEUS está financiado por la Defense Advanced Research Project Agency con la beca N00014-90-J-1851 de la Office of Naval Research, y por la National Science Foundation con la beca IRI-89-02304.

La investigación de Antonio Moreno Sandoval fue subvencionada por una beca posdoctoral MEC-Fullbright.

GRISHMAN, R. *Introducción a la Lingüística Computacional.* Visor, Madrid, 1991,

CRISHMAN, R. Information Extraction from Natural Language Text". *PROTEUS Project Memorandum* núm. 47. Department of Computer Science, New York University. Nueva York, 1991.

MARCOS, F./MORENO, A./SÁNCHEZ, F. "El proyecto EUROTRA en el marco de la investigación sobre traducción por ordenador, en *Telos,* núm. 16, diciembre-febrero 1988-89, págs. 90-99. Madrid 1989.

MORENO, A./OLMEDA, C./GRISHMAN, R./MACLEOD, C./ STERLINC, J. "PROTEUS: un sistema multilingüe de extracción de información'', en Ac*tas del VIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).* Universidad de Granada, 1992.

OLMEDA, C./MORENO, A. "El tratamiento semántico en un sistema automático de extracción de información". *PROTEUS Project Memorandum* núm. 50, Department of Computes Science, New York University, Nueva York, 1992.

Proceedings of the Message Understanding Conference-3.
San Mateo, Morgan Kauffmann, 1991.

Proceedings of the Fourth Message Understanding Conference (MUC-4). San Mateo, Morgan Kauffmann, 1992.

SAGER, N, *Natural Language Information Processing: a Computes Grammar of English and Its Applications.* Readmg, Addison-Wesley, 1981.

SUNDHEIM, B. "Overview of the Fourth Messag e Understanding Evaluation and Conference", en *Proc. Of the MUC4,* págs. 3-21. 1992.