HABÍA UNA VEZ… UNA MÁQUINA PARLANTE

Alexander Gelbukh

Revista *Ciencia y Desarrollo*, Vol. 37, N 251, 2011. ISSN 0185-0008

Platicar con Pinocho[*]

En los cuentos para los niños, los animales y las cosas aparentemente inanimadas, pero mágicas, se comportan como personas; es decir, *inteligentemente*; pueden ver, oír, pensar, actuar…, ¿cómo sabemos que un animal o una cosa son inteligentes? Porque son parlantes: hablan y entienden lo que se les dice. Desde hace miles de años el ser humano ha asociado la inteligencia con el habla.

En nuestros días la ciencia convierte cada vez más cuentos en realidad. Ya no nos sorprende la idea de volar, aunque no se use una alfombra, y ¿qué falta para que podamos conversar con Pinocho? Claro, falta que sea inteligente. La ciencia que procura construir máquinas inteligentes o intenta dotar de inteligencia a las máquinas se llama Inteligencia Artificial; y en ella se estudia cómo lograr que las máquinas puedan ver, pensar, actuar y tomar decisiones. Muchos libros se han escrito sobre cada una de estas habilidades en las máquinas y cómo mejorarlas, pero en este texto sólo quiero platicar sobre una de ellas, la que a mi juicio mejor resume y manifiesta la inteligencia, trátese de humano o máquina; además, analizaremos cómo una máquina puede procesar el lenguaje, un rasgo que hasta los últimos tiempos fue completamente exclusivo de los humanos.

Por cierto, no soy el único en pensar que la habilidad de conversar es una medida de la inteligencia; uno de los fundadores de la ciencia de la computación y de la Inteligencia Artificial, el gran científico británico Alan Turing, en 1950, propuso la famosa prueba de Turing; el planteamiento es que una máquina es tan inteligente como un humano, si puede mantener una conversación como un humano.

Por Procesamiento de Lenguaje Natural (PLN, también NLP por sus siglas en inglés) se entiende la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje. En este sentido, un perico no es un animal parlante (y no conozco ningún cuento sobre un perico inteligente): aunque produce sonidos que suenan como frases, no los entiende; no comunica información con ellos, no reacciona de manera adecuada cuando le hablan. De igual manera, una contestadora telefónica, una impresora o un procesador de palabras tampoco son dispositivos o software de procesamiento de lenguaje natural: manipulan los sonidos o letras pero no información; su reacción a las frases que procesan es simple y poco depende del significado de las frases. Al contrario, un traductor automático es un ejemplo des software de PLN: su reacción a

[*] Una versión más corta de este artículo fue publicada en la revista *Komputer Sapiens* editada por la SMIA.

las frases de entrada es compleja, se determina por el significado que comunican y refleja un gran conocimiento empleado en el proceso.

Diferentes programas pueden exhibir un distinto grado del procesamiento inteligente de lenguaje; ejemplo; un buscador de documentos puede simplemente buscar los archivos que contienen la misma cadena de letras que especificó el usuario, sin importar que ésta tenga o no significado en una lengua particular (como español o inglés), y en este caso no habría una aplicación del PLN; sin embargo, puede buscar los documentos que comunican la idea especificada por el usuario, sin importar con qué letras se presente: esa sería una excelente aplicación de PLN, ya que *la máquina* entendería la idea comunicada por el usuario, en cada documento, y sería capaz de compararlos.

Usualmente, aún la mínima capacidad para razonar sobre la información, y no sólo sobre las letras, aumenta de manera importante la utilidad de un programa; así, aunque el gran sueño de los investigadores que trabajamos en esta área es poder algún día conversar en viva voz con Pinocho, avances, incluso, muy pequeños e insignificantes en comparación con este sueño, llevan a grandes logros tecnológicos en las aplicaciones de las tecnologías de PLN.

* * *

Aparte del sueño de conversar con Pinocho (de lo cual quizás estamos menos lejos de lo que parece), el PLN tiene un gran número de aplicaciones tanto dentro de la ciencia como en la práctica.

Lingüística, hacia una ciencia empírica y exacta

Antes de que procedamos a discutir las aplicaciones prácticas del PLN, quiero mencionar que, para la ciencia sobre el lenguaje humano, el PLN es una muy poderosa herramienta de investigación, tal como un microscopio para la biología o un telescopio para la astronomía.

La lingüística estudia el rasgo más importante que nos diferencia de los animales: el lenguaje humano. Antes de los recientes avances en el procesamiento automático del lenguaje, la investigación lingüística fue, en gran medida, un asunto de introspección y especulación, lo cual no significa que no había logros importantes en esta ciencia; al revés, logró ideas, métodos y conocimientos impresionantes; pero su desarrollo (como en otras áreas de las humanidades) se detenía por falta de un criterio claro de la verdad: las explicaciones de las ideas se apoyaban mucho en ejemplos, en el sentido común y la buena voluntad del lector.

Con la llegada de los métodos computacionales, la situación cambió en dos aspectos. El primero ocurrió en el lenguaje en el que se expresan los términos y las ideas: en lugar de apelar al sentido común e intuición del interlocutor, los lingüistas contemporáneos expresan sus ideas o reglas con la claridad necesaria para que un dispositivo mecánico las aplique sin la intervención humana; y si no puede hacerlo, entonces habrá que refinar la idea. Con eso, la lingüística se convierte en una ciencia exacta.

El segundo cambio es la conversión en una ciencia empírica. El lingüista contemporáneo ya no estudia su propia intuición sobre el lenguaje, sino la naturaleza externa a él mismo: los datos, los textos escritos –que afortunadamente abundan en Internet–. Estos dos cambios nos brindan nuevas oportunidades para entender mejor uno de los milagros más misteriosos de la naturaleza: el lenguaje humano.

Manejo eficiente de nuestro mayor tesoro

El conocimiento es el mayor tesoro que posee la humanidad. Durante miles de años, la actividad más importante del hombre ha sido el producir conocimiento, guardarlo y pasarlo a las siguientes generaciones. Ahora bien, cuando se trata de dinero, sabemos cómo usarlo más eficientemente, lo guardamos de manera que podamos encontrarlo rápidamente cuando lo necesitemos, procuramos que no pierda valor con el tiempo; pero si del conocimiento se trata, lo manejamos de manera tan negligente como nunca podríamos imaginar manejar el dinero.

Almacenamos el conocimiento y lo transmitimos en forma de lenguaje humano –los textos escritos, por ejemplo–; sin embargo, usamos estos textos ineficientemente. Se puede mencionar cuatro componentes necesarios para el uso eficiente de tal conocimiento:

El primero es la digitalización de los documentos (o la obtención del texto como una secuencia de letras, no como una fotografía digital, a través del reconocimiento óptico de caracteres, OCR por sus siglas en inglés). Las bibliotecas tienen toneladas de libros en papel. Archivos como el Archivo General de la Nación, tienen muchos kilómetros de estantes con documentos de gran importancia, muchos de los cuales están en tal estado físico que, el simple hecho de tomarlos resulta problemático. El esfuerzo de digitalización requiere de gran complejidad del conocimiento y las heurísticas empleadas en el procesamiento de lenguaje natural. Es porque un lector humano, cuando lee un texto en el que ciertas letras no son muy claras, o escucha una conversación en un ambiente ruidoso, fácilmente restaura las partes faltantes porque entiende su contenido; para un programa es un gran reto. Los programas, hoy en día, son cada vez más capaces de reconocer el

texto impreso, incluso, el escrito a mano, o reconoce el lenguaje oral, gracias a sus capacidades lingüísticas, pero todavía falta mucho para la perfección.

En parecer este problema sólo existe en relación de los documentos viejos producidos en papel, pero no para los documentos nuevos. No es así. Aún en los documentos que se están creando con los medios electrónicos (como este mismo artículo), las relaciones lógicas entre sus partes no son explícitas; por ejemplo, en un artículo es difícil para un programa decidir qué es su título, autor, epígrafe, referencia bibliográfica, resumen, palabras clave, o decidir si una frase es un rótulo de una ilustración, una fórmula, un título de la sección, un comentario, un ejemplo lingüístico, etc. Todos estos elementos sólo se distinguen por su posición en la página y el tamaño de la fuente. Para una persona es obvio cuál es cuál porque entiende su contenido; entonces un programa también lo debe entender a cierto grado para procesar el documento correctamente: por ejemplo, decidir quién es su autor.

Es segundo componente es la búsqueda de la información relevante, llamada también recuperación de información: no es útil un conocimiento escrito y guardado si no se le puede hallar cuando se necesita. El mayor problema técnico de la búsqueda de la información es que la misma idea se puede expresar con muy diferentes palabras. Por ejemplo, el usuario puede expresar su interés con la frase “la derrota de Maximiliano I” y el documento más relevante para tal petición de búsqueda puede ser “la victoria de Benito Juárez”. Los dos textos no tienen ni una palabra en común, aunque un lector humano, usando cierta experiencia lingüística (derrota-victoria), así como cierto conocimiento de la historia (Maximiliano-Juárez) fácilmente detectaría la relevancia del documento para la petición.

Un progreso muy significativo se ha logrado recientemente, en busca de que los programas puedan utilizar este tipo de razonamiento para satisfacer de la mejor manera las necesidades de los usuarios en cuanto a la búsqueda de los documentos relevantes.

El tercer componente en el manejo óptimo del conocimiento es la presentación eficiente de la información contenida en los textos. Un ejemplo más directo de este tipo de tecnología es la construcción automática de resúmenes: dado un texto largo (o una colección de textos, la cual puede contener millones de documentos). Un generador automático de resúmenes trata de detectar lo más importante de los documentos, y lo presenta al lector en forma de un texto corto que él podrá leer en un tiempo razonable. A pesar de mucho esfuerzo que los expertos en el PLN han dedicado a estas tecnologías, los resultados obtenidos hasta ahora son perfectibles, aunque cada vez mejores.

Otra manera de resumir la información contenida en muchos documentos y hacerlos más manejables es agruparlos y clasificarlos; así el usuario, en lugar de manejar millones de archivos, sólo necesitará considerar, digamos, cinco grupos, en los cuales los documentos son parecidos entre sí; o bien, pueden ser diferentes personas quienes considerarán cada grupo de documentos. Un ejemplo de la aplicación de la clasificación de documentos es el enrutamiento de las quejas y peticiones de los ciudadanos a las oficinas correspondientes del gobierno a cualquier nivel (o una empresa grande).

El resumen de la información relevante contenida en un gran número de documentos puede llegar a ser tan corto como una sola palabra. Es el caso de la respuesta automática a preguntas. El usuario de un sistema de recuperación de información quiere encontrar un documento; pero, ¿para qué lo quiere encontrar? Es muy probable que en realidad no necesite el texto específico, pero tiene una duda e intenta aclararla leyendo los documentos. Las tecnologías de respuesta automática a preguntas lo hacen directamente: a la petición "¿Dónde nació Juárez?" la respuesta será "en Guelatao", nótese que un programa de recuperación de información entregaría al usuario la biografía completa de Juárez para que busque el dato. Los sistemas de respuesta automática a preguntas están basados en un razonamiento complejo que, a veces, requiere de la profunda comprensión del significado del texto; ejemplo: el documento que contiene la respuesta relevante podría ser "... *llegamos a Guelatao, el pueblo natal del Benemérito de Las Américas* ..."; no es una tarea trivial para un programa –aunque sí lo es para un humano– deducir de este texto que Juárez nació en Guelatao.

Otras tecnologías que usan directamente el contenido de los textos incluyen la minería de texto (encontrar las opiniones prevalecientes expresadas en los textos, las tendencias de cambio de estas opiniones o las relaciones inesperadas entre los eventos descritos en los textos), la extracción de información (llenar bases de datos sobre un tema específico, leyendo los textos) y sistemas de soporte a la toma de decisiones (buscar, sintetizar y presentar de manera eficiente la información relevante para un directivo).

Finalmente, el cuarto paso en el manejo eficiente de la información va más allá de entregar al usuario un texto para su lectura, ya sea completo o resumido. Se trata más bien del uso de la información contenida en los textos por el mismo software para resolver tareas más complejas. Por ejemplo, la máquina puede aprender automáticamente el conocimiento necesario de los textos disponibles en Internet, tales como los artículos científicos o los libros de texto. Las aplicaciones de este tipo están actualmente en la fase experimental, aunque en el futuro inevitablemente se convertirán en la manera principal del manejo de conocimiento.

Entender el lenguaje ajeno es la paz

Parafraseando la célebre frase, se puede decir que *el entender el lenguaje de otros es la paz*. Los individuos, las naciones y los pueblos, se unen gracias a un lenguaje común (como es el caso de los pueblos de la América Latina); así como se dividen (política, económica, social y culturalmente) por las fronteras, no tanto políticas, sino lingüísticas (como también se puede observar en el mapa de nuestro continente). Los individuos, como las naciones, los pueblos o grupos pueden sentirse excluidos (económica, social y culturalmente) por la frontera lingüística, la cual les dificulta el acceso a la información producida por la humanidad.

A los esfuerzos por combatir los efectos negativos de la división lingüística en el mundo y en nuestro país, la ciencia del procesamiento de lenguaje natural aporta las tecnologías de la traducción automática. Con esta tecnología, el usuario puede leer en su propia lengua un texto originalmente escrito en otro; puede escribir sus ideas para los lectores que hablan otro idioma, o hasta puede conversar (ya sea a través de los mensajes instantáneos o en viva voz) con un interlocutor hablante de una lengua diferente.

La calidad de la traducción automática se mejoró dramáticamente en la última década. Hace unos diez años, los sistemas experimentales fueron usados principalmente para acelerar un poco el trabajo de los traductores profesionales; los textos generados requerían mucha corrección manual (a excepción de los sistemas capaces traducir los textos de una temática muy específica, tal como los pronósticos del estado de tiempo). En cambio, hoy en día el traductor de Google (www.google.com.mx/language_tools?hl=es) produce el resultado completamente legible y usable para que podamos, sin ayuda externa, leer las páginas de Internet en chino, árabe, ruso y muchas otras lenguas, sin mencionar el inglés.

Mientras que el texto producido por este traductor (u otros automáticos) es indudablemente útil y sirve de gran ayuda, no hace falta aclarar que es muy superable, en dos de los aspectos más importantes que actualmente son aún deficientes:

- Primero, la calidad del texto que se produce. En muchas ocasiones suena como escrito por un extranjero que no habla bien el español (en nuestro caso) y, en otras, de plano nos reprobarían en la primaria si escribiéramos así. Aunque el mejorar este aspecto requiere de mucho esfuerzo, es más manejable (y el progreso en esto se nota más) que el segundo problema y, por otro lado, aunque a veces el texto se ve raro, no presenta tanta molestia en la práctica.
- El segundo problema es la traducción incorrecta. Este problema se nota mucho menos que el primero (y entre más necesita el usuario la ayuda del traductor, menos va a notar sus errores), pero puede causar consecuencias mucho más graves por los posibles malos entendidos e

información falsa. Además, es mucho más difícil corregir este tipo de problemas —es decir, desarrollar un software para la traducción automática que evite al máximo las alteraciones del significado en la traducción.

Esta tarea necesita toda la fuerza de la ciencia del procesamiento de lenguaje natural y, en muchos casos, requiere que el programa entienda el texto en un nivel lo suficientemente profundo para poder razonar sobre él. Con justa razón, la traducción automática, desde el mismo comienzo de esta ciencia, fue su principal motivación y la fuente de inspiración y retos.

Sin embargo, vale la pena; una vez resueltos los problemas técnicos, viviremos en un mundo sin fronteras lingüísticas, sin limitaciones impuestas a uno por no hablar inglés, chino o español; para hablar con un vecino de continente, simplemente prenderíamos el celular que se encargaría de decirle en inglés lo que le estamos diciendo en español, y de igual manera traducir su respuesta. O bien, el navegador nos mostraría todo el Internet en español, sin importar en qué idioma escribió cada página su autor.

Era informática para todos

Vivimos en una era informática de libre acceso a la información, de trabajo intelectual eficiente, al ser asistido por computadora. En esta era vivo yo, mis colegas ingenieros, mis estudiantes y seguramente usted, querido lector. Pero cuando hablo con algunos de mis amigos médicos, abogados, músicos, historiadores, chóferes, campesinos, amas de casa…, me sorprende cuán poquitos vivimos en esta era. Lo que escucho de ellos es “pues, la computadora y estas cosas... es que ¡no soy bueno en esto!” Pero eso no es cierto, la que es mala es la computadora.

Las computadoras fueron creadas para resolver nuestros problemas y no para crearnos más (como la necesidad de aprender a programarlas). Deben ser nuestras ayudantes naturales y fáciles de usar; deben aprender nuestro lenguaje y no obligarnos a aprender el suyo.

En breve, las máquinas (hablo aquí más de los robots que de las computadoras de escritorio) serán físicamente capaces de ser nuestras sirvientes y ayudantes en tareas cotidianas. En 2006, el gobierno de Corea del Sur anunció un programa según el cual cada familia coreana, en el año 2020, tendrá un robot ayudante de la casa.[1] En pocos meses Bill Gates, el líder de Microsoft, aseguró que pronto habrá un robot en cada hogar, y no sólo en Corea.[2]

Para que un robot se convierta en un verdadero ayudante de casa, debe entender nuestro lenguaje: al menos, qué instrucciones se le dan, y responder cuando *necesite* decir algo. Esto significará el inicio de la era informática para todos, no sólo para los programadores e ingenieros.

Entre muchos problemas técnicos en este camino mencionaré aquí cuatro, ninguno de los cuales es inherente a la tarea de las interfaces en lenguaje natural, pero aquí la necesidad de su solución es más evidente y los retos son más difíciles.

- El procesamiento de habla. Varias veces en este artículo dije que los programas de PLN procesan, clasifican y analizan el texto, pero no debe ser así; no hablamos en texto, hablamos en voz. Para lograr una interfaz eficiente, la máquina debe entender la voz, por ejemplo, transformarla primero a texto y luego analizar este texto, si así le conviene al desarrollador.
- La conducción del diálogo, que presenta algunos retos distintos de los que presenta un texto normal; por ejemplo, en el diálogo se usan mucho los pronombres (las palabras como “él”) o las oraciones incompletas, o hasta recortadas a una sola palabra (como “ajá”, “pues”). Además, hay ciertas reglas de conducta en cuanto al cambio de turnos: ¿cuándo dejo de escuchar y empiezo a hablar? ¿Cuánto puedo hablar sin ser interrumpido?
- La generación de lenguaje: hablar o escribir a diferencia de escuchar o leer; componer a diferencia de analizar. ¡Cuántas veces tenemos mucho que decir y lo queremos decir todo a la vez!, pero esto no se puede; hay que decidir cuál idea voy a expresar en la primera oración y cuál en la segunda (y peor aún, dividir todo lo que pensamos en pedacitos, con los que se puede formular en una sola oración), cuál palabra va primero y cuál luego; cuál palabra hay que usar para expresar la misma idea en diferentes contextos (digamos, para decir “muy” de la voz o temperatura, decimos “alta”, pero de café, decimos “cargado” y del trabajo, “duro”).
- El cuarto problema consiste en relacionar las palabras con las acciones, objetos y circunstancias que tienen relación con la conversación. El robot debe poder reaccionar adecuadamente a las frases como “éste no me gusta”, “ve allá” y “cámbiamelo por otro”, relacionando el objeto y la dirección con el movimiento del dedo del usuario y adivinando en qué debe diferir el otro (¿más frío?, ¿más caliente?, ¿con limón?).

Igual ocurre en el caso de otras aplicaciones, mientras los investigadores nos están acercando a lo que hoy se ve como ciencia ficción, existen usos de tal tecnología completamente prácticos y factibles hoy mismo. En cuanto a las interfaces humano-computadora, una aplicación práctica son las interfaces con las bases de datos. Normalmente las preguntas son aún bastante sencillas (¿qué porcentaje de los alumnos del tercer semestre obtuvieron una calificación mayor de nueve de dos o más materias?) implican programación en un lenguaje especializado de consulta a bases de datos llamado SQL (por sus siglas en inglés: Structured Query Language). Mucho esfuerzo se ha dedicado durante décadas a que los programas puedan directamente entender las preguntas en su forma

natural, proporcionando así el acceso a la información a los usuarios comunes sin la necesidad de un programador intermediario.

Un ejemplo de la aplicación práctica del reconocimiento de habla son los sistemas de dictado, entre los cuales, probablemente el más conocido es Dragon Naturally Speaking (Dragón Hablando Naturalmente) de la empresa IBM. Con tal sistema se puede dictar los textos (como este artículo) a la computadora, en lugar de escribirlos con el teclado. La miniaturización de los sistemas electrónicos llevará al crecimiento de la importancia de la entrada de los datos o comandos con voz: pronto será la única (y muy natural) manera de interactuar con un celular o un reloj de pulsera inteligente.

Para ejemplificar las aplicaciones de los sistemas de diálogo se puede mencionar los sistemas de venta de boletos de tren o avión por teléfono, capaces de conducir un diálogo simple sobre las preferencias de viaje del usuario.

Otras aplicaciones

Además de los tres grupos de aplicaciones ya mencionados –el manejo de conocimiento, la traducción automática y las interfaces humano-computadora– el PLN constituye la parte crucial de diversos tipos de sistemas relacionados con el uso de lenguaje humano; mencionemos aquí sólo algunos.

Los sistemas de soporte para la composición de textos proporcionan varios tipos de ayuda al usuario en escribir los documentos: formatean el texto usando guiones, verifican la ortografía, la gramática y el estilo, completan las palabras o frases que empieza a escribir el usuario (lo que es muy útil en los celulares), proporcionan las traducciones, sinónimos y explicaciones de las palabras o sugieren palabras según su descripción.[3] Las tareas de este tipo pueden variar en complejidad desde muy simples (tales como la división de las palabras con guiones) hasta muy complejas; por ejemplo, la verificación lógica y factual del texto (en la frase “al salir de Francia, Juan visitó Londres, su capital” un buen programa encontraría un error lógico y un error factual).

Las aplicaciones del PLN en la educación incluyen la evaluación automatizada de las respuestas o composiciones de los estudiantes en cuanto a su estilo, lenguaje o exactitud de las respuestas. En la educación asistida por computadora, los métodos del PLN ayudan a componer los cursos y a proporcionar al estudiante la información requerida (una tarea similar a la recuperación de información y la respuesta a preguntas).

En la medicina, son particularmente útiles las aplicaciones de minería de texto y búsqueda en las historias clínicas de los pacientes, además de los sistemas especializados de búsqueda y minería

de texto para los médicos. Debido a la enorme cantidad de datos experimentales reportados, por ejemplo, en la investigación de la interacción de los genes y las proteínas, resulta necesario el procesamiento automático de tales publicaciones, ya que una persona no puede leerlas, no sólo todas, sino las más relevantes para su trabajo.

El término *lingüística forense* se refiere a las diversas aplicaciones de los métodos lingüísticos, y sobre todo computacionales, en las investigaciones criminalísticas y peritaje. Estos métodos incluyen la identificación de la autoría de los textos o búsqueda de los fragmentos sospechosos en los mensajes o conversaciones grabadas. Dos áreas son muy afines a la lingüística forense: una es la identificación de plagio (tanto en obras literarias o publicaciones científicas como en las composiciones de los estudiantes); la otra es la esteganografía lingüística –los métodos para ocultar mensajes secretos en textos o habla, así como los métodos para detectar tales mensajes ocultos–.

Resulta interesante que las ideas y técnicas desarrolladas originalmente para el análisis de lenguaje resultan aplicables en las áreas muy lejanas del lenguaje humano. Un ejemplo obvio es la teoría de compiladores y los lenguajes de programación, la creciente complejidad de los cuales cada vez los aproxima a los lenguajes humanos. Perl (lenguaje de programación diseñado por el lingüista Larry Wall en 1987) es un ejemplo de un lenguaje computacional intencionalmente diseñado para aprovechar algunos rasgos de los lenguajes humanos, tales como la ambigüedad.

La genómica y la biología molecular comparten muchos ideas y métodos con el PLN, lo que no es tan sorprendente como parece, ya que en ambos casos se trata de la codificación de la información compleja en una cadena de símbolos, la cual en el caso de la genómica es la molécula de ADN, ARN o las moléculas de las proteínas.

Finalmente, y por las razones similares, los métodos de PLN se emplean en el análisis y la generación automática de la música. Resulta que las estructuras repetitivas musicales se pueden describir bien con las así llamadas gramáticas formales desarrolladas originalmente para la descripción de los fenómenos lingüísticos.

Problemas del PLN

En general, los mexicanos, desde la niñez más temprana, hablamos español sin problema; las máquinas, que son más rápidas, deberían entenderlo aún más fácilmente. ¿Cuál es entonces el problema técnico en la implementación de los sistemas del PLN?

En el pasado reciente el reconocimiento de las estructuras de las palabras y las frases fue un problema originado en la baja velocidad de los procesadores y la poca memoria disponible. Hoy en día, los problemas de este tipo parecen ser, en su mayor parte, resueltos, o al menos, nos es

suficientemente claro cómo resolverlos; no obstante, dos problemas de fondo siguen siendo el mayor obstáculo para que los programas puedan entender el lenguaje natural.

El primer problema es la ambigüedad. Este fenómeno consiste en que la misma expresión se puede interpretar de diferentes maneras. Por ejemplo, la palabra "gato" se puede entender como un animal o una herramienta. En la oración "Juan come arroz con palillos" no es completamente claro si Juan come los palillos junto con arroz o los usa para comer el arroz, compárese con la oración "Juan come arroz con leche". En la oración "Juan tomó la torta de la mesa y la comió" no es tan obvio –al menos para una máquina que no conoce todos los significados de las palabras– qué es lo que comió Juan, la torta o la mesa (compárese con la oración "Juan tomó la torta de la mesa y la limpió con un trapo" —¿qué es lo que limpió Juan?). Usualmente, no es difícil para el programa encontrar una interpretación sobre el texto, o mejor, encontrar todas posibles interpretaciones; lo difícil es elegir la correcta en cada caso.

Es allí donde nos encontramos con el segundo problema del procesamiento de lenguaje: el conocimiento necesario para resolver la ambigüedad. Básicamente lo que el programa necesita conocer es la diferencia entre el gato animal y el gato herramienta; si la gente normalmente usa los palillos (o la leche) para comer arroz o los (la) come junto con él; qué es lo que se come normalmente, las tortas o las mesas. Aunque cada uno de tales hechos parece trivial, son tantos hechos diferentes que, en la actualidad, la construcción de una base de datos enorme que los contenía todos parece ser una tarea, al menos, muy difícil.

Con esto, no es tan sorprendente el hecho de que los programas del PLN resultan ser complejos y su desempeño, en muchos casos, muy mejorable. Lo que es realmente sorprendente es cómo los niños aprenden a solucionar este tipo de problemas (y todos los demás relacionados con la comprensión de lenguaje) en un tiempo récord, básicamente durante el primer año de su vida. La respuesta corta es: no se sabe esto todavía; hay muchas teorías, todas interesantes, que procuran explicar este fenómeno, pero debemos confesar que esto sigue siendo uno de los mayores misterios de la naturaleza. Esperamos, sin embargo, que la ciencia del procesamiento de lenguaje natural contribuirá, finalmente, al descubrimiento de este misterio.

Métodos del PLN

Bueno, y si son tan difíciles los problemas de la comprensión del lenguaje, ¿es factible resolverlos? Sí, es factible. La ciencia que estudia el procesamiento automático del lenguaje natural se llama lingüística computacional, que en un principio sólo se denominó *lingüística*, pues se consideraba que su desarrollo consistiría en que los lingüistas, a través de la introspección e intuición, escribirían los

diccionarios y las reglas cada vez más exactos y detallados, los cuales cada vez más, nos acercarían al objetivo: dotar a la computadora de la capacidad de entender el lenguaje humano; camino que resultó muy difícil y laborioso, y los avances, aunque impresionantes, fueron lentos y esporádicos.

Todo cambió con la llegada de Internet; entonces, volúmenes gigantescos de textos se hicieron disponibles para los investigadores, es decir, los ejemplos del objeto de nuestro estudio: el lenguaje. La tarea, entonces, dejó de requerir la introspección e intuición, y en cambio, requirió el estudio estadístico directo de los datos concretos disponibles inmediatamente. La lingüística computacional –al menos en la primera etapa de su desarrollo actual– dejó de ser una rama de la lingüística y se convirtió en otra rama de la ciencia: el *aprendizaje automático*, la cual forma parte tanto de la inteligencia artificial como de la estadística.

El aprendizaje automático tiene como fin el descubrimiento totalmente automático de las regularidades y las relaciones establecidas entre los datos. Usualmente los datos a los cuales se aplican los algoritmos de aprendizaje son numéricos. Sin embargo, se puede plantear la tarea de aprendizaje cuando estos datos son una secuencia larga de letras, a saber, un texto, y las regularidades que se buscan son representan el conocimiento lingüístico sobre el lenguaje en el cual están escritos estos textos. La investigación de las técnicas de este tipo de aprendizaje constituye una gran parte de la lingüística computacional contemporánea. Tal aprendizaje automático, a partir de una muy larga secuencia de letras, es como un niño aprende su lenguaje natal: nadie le enseña las reglas, las gramáticas y los diccionarios, sino su cerebro analiza estadísticamente los sonidos del lenguaje (y, a diferencia de los programas de PLN actualmente usados, su relación con el medio ambiente) y aprende a usarlas adecuadamente.

El programa tiene acceso a una enorme cantidad de textos (los datos de entrenamiento), en los cuales se puede encontrar automáticamente los ejemplos útiles para el razonamiento, y consiste en elegir una de las dos acepciones de la palabra en el diccionario monolingüe. Una vez decidido qué gato tiene Juan, el programa es capaz (usando un diccionario bilingüe) de traducir la frase correctamente al inglés. Sin razonamiento de este tipo, la traducción muy probablemente saldría totalmente incorrecta: “John uses a cat to repair his car”, la cual básicamente dice que Juan usa un minino para reparar su carro.

Expertos en el PLN

No es el propósito de este texto introductorio el explicar al lector los pormenores técnicos, sino despertar su interés por las tecnologías del PLN. Ahora bien, suponiendo que haya logrado este propósito, sólo me queda decir dónde el lector podrá encontrar a los expertos en el área. Si es usted

directivo o empresario y encontró en este artículo algo que le puede servir, me preguntaría quién le puede dar el servicio, y si es usted estudiante, quizá potencial –nunca es tarde estudiar–, me preguntaría dónde puede obtener más información. Una buena fuente sobre el tema son nuestros libros,[4, 5, 6] disponibles en texto completo desde la página www.Gelbukh.com; sin embargo, la mejor manera de aprender es colaborar con los expertos en el área.

Para empezar, el mapa de la figura 1 indica los países de origen de los autores de las ponencias sometidas en los últimos cinco años al congreso CICLing (www.CICLing.org), un congreso internacional sobre el PLN que organiza anualmente el Laboratorio de Procesamiento de Lenguaje Natural del Centro de Investigación en Computación del IPN. El área de cada círculo simboliza el número de los autores provenientes del país correspondiente, y se nota que las grandes áreas donde se cultiva el PLN son Europa, los EUA y China (esta interpretación de los datos es muy simplificada, ya que el número de ponencias puede depender mucho de las políticas institucionales que fomentan o no, su presentación en los congresos).

En particular, en España existen varios grupos muy buenos que cultivan el PLN: la Universidad de Alicante, la Universidad Politécnica de Barcelona, la Universidad Politécnica de Valencia, la Universidad Nacional de Educación a Distancia, por mencionar sólo algunos.

En la América Latina, México mantiene el liderazgo (seguido por Brasil; muy pocos son los grupos en otros países latinos). Las instituciones que cultivan esta ciencia en México incluyen el INAOE, el IPN, la UNAM, la UAM, la BUAP, la UAEM, el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), la Universidad Autónoma del Carmen, entre otras.

La comunidad nacional del PLN organiza anualmente, al menos dos congresos nacionales y las ponencias se presentan en español: se trata del Coloquio de Lingüística Computacional en la UNAM y el Taller de Tecnologías del Lenguaje Humano, organizado por el INAOE. Además, el IPN organiza anualmente el congreso internacional CICLing, ya mencionado, aunque no siempre tiene sede en México.

Finalmente, como he mencionado al inicio del artículo, la ciencia del procesamiento de lenguaje natural está estrechamente integrada con muchas otras ciencias, que en su totalidad forman el área de la inteligencia artificial: la ciencia que estudia cómo construir las máquinas más inteligentes. Los expertos mexicanos en la inteligencia artificial y todas sus ramas están unidos alrededor de la Sociedad Mexicana de Inteligencia Artificial (SMIA, www.SMIA.org.mx), la cual organiza anualmente dos magnos eventos de la vida científica en México: el congreso internacional *Mexican International Conference on Artificial Intelligence* (MICAI, www.MICAI.org) y el Congreso Mexicano de Inteligencia Artificial (COMIA). La SMIA y sus congresos son los sitios más adecuados de

búsqueda para el lector mexicano interesado en la amplia gama de problemas y soluciones en el camino al hacer realidad el cuento de una máquina parlante –una máquina inteligente–.

Referencias

1. A Robot in Every Home by 2020, South Korea Says. http://news.nationalgeographic.com/news/2006/09/060906-robots.html, visitado el 11 de febrero de 2010.
2. B. Gates. A Robot in Every Home. Scientific American, 2007; http://www.scientificamerican.com/article.cfm?id=a-robot-in-every-home, visitado el 11 de febrero de 2010.
3. Gerado Sierra. Búsqueda de palabras a partir de las definiciones en los diccionarios de lengua automatizados. Simposios Internacionales de Comunicación Social, Simposio 7, Actas 2, Santiago de Cuba, 2001.
4. A. Bolshakov, A. Gelbukh. Computational linguistics: models, resources, applications. IPN – UNAM – Fondo de Cultura Económica, 2004.
5. Gelbukh, G. Sidorov. Procesamiento automático del español con enfoque en recursos léxicos grandes. Segunda edición, ampliada y revisada. IPN, 2010.
6. S. N. Galicia Haro, A. Gelbukh. Investigaciones en análisis sintáctico para el español. IPN, 2007.

Anexo

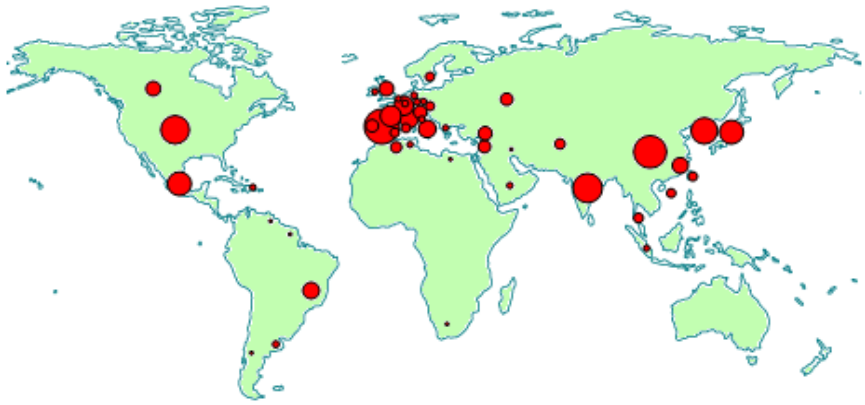

Figura 1. Lugares donde se cultiva más la lingüística computacional

Alexander Gelbukh es maestro en matemáticas por la Universidad Estatal Lomonósov de Moscú, y doctor en ciencias de la computación por el Instituto de Información Científica y Técnica de Toda Rusia. Desde 1997 es jefe del Laboratorio de Procesamiento de Lenguaje Natural del Centro de Investigación en Computación del IPN. Actualmente es, además, investigador visitante en la Universidad Waseda, Tokio, Japón. Es miembro de la Academia Mexicana de Ciencias, investigador nacional, nivel 2 y Vicepresidente de la Mesa Directiva de la Sociedad Mexicana de Inteligencia Artificial (SMIA). Es autor, coautor o editor de alrededor de 450 publicaciones y coautor de tres libros en las áreas del procesamiento de lenguaje natural e inteligencia artificial.