# Procesamiento de lenguaje natural

por **Alexander Gelbukh**


## Un cuento de una máquina parlante

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

En nuestros días la ciencia convierte cada vez más cuentos en la realidad. Ya no nos sorprende una alfombra voladora (aunque no parezca a una alfombra) y ¿que falta para que podamos conversar con el Pinocho? En los números anteriores de nuestra revista le hemos contado al lector sobre cómo las máquinas pueden ver, pensar, actuar, tomar decisiones. En este número vamos a platicar sobre cómo una máquina puede procesar el lenguaje —un rasgo que hasta los últimos tiempos fue completamente exclusivo de los humanos.

Por procesamiento de lenguaje natural (PLN, denominado 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 muchos cuentos sobre los pericos parlantes); así, una contestadora telefónica común, una impresora o un procesador de palabras como Microsoft Word tampoco son dispositivos o software de procesamiento de lenguaje natural, mientras que un traductor automático sin duda lo es.

Diferentes programas pueden exhibir diferente grado del procesamiento inteligente de lenguaje. Por ejemplo, un buscador de documentos puede simplemente buscar los documentos que contienen la misma cadena de letras que el especificó el usuario, sin importar que esta cadena de letras tenga o no un significado en un lenguaje particular (como el español o el inglés). En este caso no sería una aplicación del PLN. Sin embargo, puede buscar los documentos que comunican la idea que especificó el usuario, sin importar con qué letras la comunican —en este caso, sin duda, sería una excelente aplicación de PLN, ya que entendería la idea comunicada en la petición del usuario, la idea comunicada en cada uno de los documentos, y sería capaz de compararlas.

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

## Aplicaciones

Aparte del sueño de conversar con Pinocho (de lo cual quizá estamos menos lejos de que parezca), el PLN tiene un gran número de aplicaciones tanto dentro de la misma 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, como un microscopio para la biología o un telescopio para la astronomía.

La lingüística estudia el rasgo más importante que nos difiere de los animales: el lenguaje humano. Antes de los recientes avances en el procesamiento automático de lenguaje, la investigación lingüística fue en gran medida un asunto de introspección y especulación. No quiere decir esto que no había logros importantes en esta ciencia; al revés, logró ideas, métodos y conocimientos impresionantes. Sin embargo, el desarrollo de esta ciencia (como algunas otras áreas de las humanidades) se detenía por no haber un criterio claro de verdad: los científicos de diferentes escuelas discutían sobre lo que a uno le parece cierto y lo él cree, pero en otro no cree. Las explicaciones de las ideas se poyaban mucho en los ejemplos, en el sentido común y buena voluntad del lector, con la esperanza de que al lector le será obvio lo que le dicen.

Con la llegada de los métodos computacionales la situación se cambió en dos aspectos. El primero es un cambio de lenguaje en el cual 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 pueda aplicar sin la intervención humana. Y si no lo puede, entonces habrá que refinar la idea. Con eso, la lingüística se convierte en una ciencia exacta.

El segundo cambio es que se convierte en una ciencia empírica. El lingüista contemporáneo ya no estudia su propia intuición sobre el lenguaje sino estudia la naturaleza externa a él mismo: los datos, los textos escritos —los 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 del nuestro mayor tesoro: Búsqueda y manejo de conocimiento

El conocimiento es el mayor tesoro que posee la humanidad. Durante miles de años la actividad más importante del hombre ha sido el producir el conocimiento, guardarlo y pasarlo a las siguientes generaciones. Ahora bien, cuando se trata de dinero, sabemos cómo usarlo más eficiente, lo guardamos de tal manera para encontrarlo rápidamente

cuando lo necesitamos, procuramos que no pierda valor con el tiempo. Sin embargo, cuando se trata de nuestro mayor tesoro —el conocimiento—, lo manejamos de la manera tan negligente como nunca podemos imaginar manejar el dinero.

El conocimiento lo almacenamos y transmitimos en forma de lenguaje humano —los textos escritos, por ejemplo, en español o inglés. Sin embargo, usamos estos textos muy ineficientemente. Se puede mencionar cuatro componentes necesarios para el uso eficiente de tal conocimiento.

El primero es la digitalización de los documentos. Las bibliotecas tienen toneladas de libros en papel. Los archivos, tales como el Archivo General de la Nación, tienen muchos kilómetros de estantes llenos con documentos de gran importancia, muchos de los cuales están en tal estado físico que simplemente tomarlos en la mano es problemático. Aún los documentos que se están creando con los medios electrónicos (como este mismo artículo) luego se publican sólo en papel, o su formato inicial se pierde, como es el caso de las fórmulas matemáticas.

El esfuerzo de digitalización se requiere de gran fuerza de procesamiento de lenguaje natural (por digitalización aquí se entiende la obtención del texto como una secuencia de letras, no la obtención de una fotografía digital). Un lector humano, cuando lee un texto donde ciertas letras no son muy claras o cuando escucha una conversación en un ambiente ruidoso, fácilmente restaura las partes faltantes porque entiende su contenido. Los programas hoy en día son cada vez más capaces de reconocer el texto impreso o hasta escrito a mano o reconocer el habla, gracias a sus capacidades lingüísticas.

Es segundo componente del manejo eficiente del conocimiento es la búsqueda de la información relevante, llamada también la recuperación de información: no es útil un conocimiento escrito y guardado si no se puede encontrarlo 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 ninguna palabra en común, aunque un lector humano, usando cierta experiencia lingüística (derrota—victoria), así como cierto conocimiento del mundo (Maximiliano—Juárez) fácilmente detectaría la relevancia del documento para la petición.

Recientemente el progreso muy significativo se ha logrado para 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 eficiente 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 que comunican estos documentos y presentarlo 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 muy mejorables, 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 tener que 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 los documentos es el enrutamiento de las quejas y peticiones de los ciudadanos a las oficinas correspondientes del gobierno o alcaldía (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 documento sino 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 le entregaría al usuario la biografía completa de Juárez para que lo busque allá. 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: por ejemplo, el documento que contiene la respuesta relevante podría se “*... 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 cuatro 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: Traducción automática

Parafraseando la frase célebre se puede decir que el entender el lenguaje ajeno es la paz. Los individuos, como las naciones y los pueblos, se unen gracias a su 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 para combatir estos 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 propio lenguaje un texto originalmente escrito en otro lenguaje, puede escribir sus ideas para los lectores que hablan otro lenguaje, o hasta puede conversar (ya sea a través de los mensajes instantáneos o en viva voz) con un interlocutor que habla otro lenguaje.

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 y 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 traductores automáticos) es indudablemente útil y sirve de gran ayuda, no hace falta aclarar que es muy mejorable. Son dos aspectos que más importantes en que son actualmente deficientes estos sistemas.

Primero, la calidad del texto que producen. En muchas ocasiones suena como escrito por un extranjero que no habla bien el español, 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 a lo 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 pueda entender el texto a 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 el inglés (o el chino, o el español) y sin tanta división cultural y social derivada de las fronteras lingüísticas. 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 va a mostrar todo el Internet en español, sin importar en qué idioma escribió cada página su autor.

## Era informática para todos: Interfaces humano-computadora

Vivimos en una era informática. En una era de libre acceso a la información. En una era de trabajo intelectual eficiente por ser asistido por la computadora.

Digo —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 cuántos 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 no es cierto. Son buenos. La que es mala es la computadora.

Las computadoras fueron creadas para resolver nuestros problemas y no para crearnos más problemas (digamos, 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 bien de los robots y no de las computadoras de escritorio) estarán físicamente capaces de ser nuestras sirvientes y ayudantes en nuestras tareas cotidianas. En 2006 el gobierno de Corea del Sur anunció su programa según el cual cada familia coreana en el año 2020 tendrá un robot ayudante de la casa [1], tal como en los signos pasados era común tener los sirvientes. 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, tiene que entender nuestro lenguaje: al menos entender qué le dicen hacer y responder cuando necesita 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 cuatro 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 primero es el procesamiento de habla. Varias veces en este artículo dije que los programas de PLN procesan, clasifican, 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.

El segundo es la conducción del diálogo. Un diálogo presenta algunos retos distintos de los que presenta un texto normal, un monólogo. Por ejemplo, en el diálogo se usan mucho los pronombres (las palabras como “el”) 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?

El tercer problema es 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 que pensamos en pedacitos 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").

Finalmente, el cuarto problema es relacionar las palabras con las acciones, objetos y circunstancias relacionados a 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 frio? ¿más caliente? ¿con limón?).

Igual como 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 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 en 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 avió por teléfono, capaces de conducir un diálogo simple sobre las preferencias de viaje del usuario.

## Y mucho, mucho más: Otras aplicaciones

Además de los tres grupos de aplicaciones ya mencionados —el manejo de conocimiento, la traducción automática y las interfaces humano-computadora— es 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 proporcionarle al estudiante la información requerida (una tarea similar a la recuperación de información y la respuesta a preguntas).

En la medicina, particularmente útiles son 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 enrome cantidad de los datos experimentales reportados, por ejemplo, en la investigación de la interacción de los genes y los proteínas, resulta necesario el procesamiento automático de tales publicaciones ya que una persona ya no puede leerlas no solo todas sino las más relevantes para su trabajo.

El término "lingüística forense" 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 literarios 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 computaciones, la creciente complejidad de los cuales cada vez los aproxima a los lenguajes humanos. Perl es un ejemplo de un lenguaje computacional que fue intencionalmente diseñado para aprovechar algunos rasgos de los lenguajes humanos, tales como la ambigüedad (a propósito, su autor es un lingüista).

Le genómica y la biología molecular comparten muchos ideas y métodos con el PLN. Eso 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 DNA, RNA 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

Todos nosotros desde la niñez más temprana hablamos en español sin ningún problema. Las máquinas, que son más rápidas, deberían de entenderlo aún más fácil. ¿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 debido a la baja velocidad de los procesadores y poca memoria disponible. Hoy en día los problemas de este tipo parece ser en su mayor parte resueltos, o al menos es lo suficientemente claro cómo resolverlos. Sin embargo, 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 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 claro qué comió Juan, la torta o la mesa (compárese con la oración “Juan tomó la torta de la mesa y la limpió con un trapo”). Usualmente no es difícil para el programa encontrar una interpretación para el texto, o más bien, encontrarlas todas. Lo difícil es elegir la correcta.

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 saber es cuál 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 es 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 record —básicamente durante el primer año de su vida. La respuesta corta es: no se sabe esto todavía. Hay muchas teorías muy interesantes que tratan de explicar este fenómeno, pero hay que 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 en que finalmente se descubra este misterio.

## Métodos

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. Este nombre fue inventado en los tiempos cuando eso era: lingüística, pero computacional. 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: donar la computadora con la capacidad de entender el lenguaje humano. No sorprende que resultara muy difícil y laborioso este camino, y los avances, aunque impresionantes, fueron lentos y esporádicos.

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

El aprendizaje automático tiene como fin el descubrimiento totalmente automático de las regularidades y las relaciones en los datos. Usualmente los datos a los cuales se aplican los algoritmos de aprendizaje son numéricos, pero entonces en este sentido esta parte de la lingüística computacional se puede considerar como el aprendizaje automático sobre tal tipo de datos especial —los textos en un lenguaje. Es así 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.

Un ejemplo extremadamente simplificado de tal proceso de razonamiento basado en los datos observados y no en las reglas definidas por el lingüista se muestra en el recuadro: ¿qué gato tiene Juan? El programa tiene acceso a una enorme cantidad de textos (llamada los datos de entrenamiento), en los cuales encuentra automáticamente los ejemplos útiles para el razonamiento (el cual 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) traducir la frase correctamente a inglés. Sin el 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

No es el propósito de esta introducción corta el explicar al lector los pormenores técnicos, sino más bien despertar su

interés por las tecnologías del PLN. Ahora bien, suponiendo que logré este propósito, sólo me queda decir dónde el lector podrá encontrar los expertos en el área. Si es usted un directivo o un 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 donde puede obtener más información.

Para empezar, el mapa en la ilustración 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 Instituto Politécnico Nacional. El área de cada círculo simboliza el número de los autores provenientes del país correspondiente. Se nota que las grandes áreas donde se cultiva el PLN son Europa, los EE.UU. y China (hay que aclarar que esta interpretación de los datos es muy simplificada, ya que el número de las ponencias puede depender mucho de las políticas institucionales que fomentan o no, la presentación de las ponencias en los congresos de este tipo).

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 y varios otros, por mencionar sólo algunos.

En la América Latina, México mantiene el liderazgo (seguido por Brasil; muy pocos son los grupos en otros países latinos). Este número de la revista presenta (con una excepción) los artículos escritos por los líderes de los grupos principales de nuestro país. Las instituciones que cultivan esta ciencia incluyen el INAOE, la UNAM (la cual tiene al menos tres grupos que trabajan en PLN; dos de éstos están representados en este número), la UAM, la BUAP, la UAEM, el CENIDET, la Universidad Autónoma del Carmen y el IPN, entre otras. La comunidad nacional del PLN está unida alrededor de varias organizaciones y redes de colaboración. La ilustración 2 introduce al lector a los fundadores de la AMPLN, la Asociación Mexicana de Procesamiento de Lenguaje Natural (www. AMPLN.org).

Esta comunidad organiza anualmente al menos dos congresos nacionales donde las ponencias se presentan en español: el 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 antes mencionado, aunque no siempre lo organiza en México. Para mayor información se puede recomendar al lector contactar a cualquiera de los autores de los artículos presentados en este número, o bien directamente a la AMPLN. También le puedo recomendar nuestros libros [4, 5, 6], disponibles en texto completo desde la página www.Gelbukh.com.

## 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.

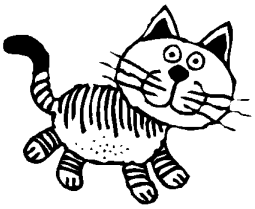
[4] I. A. Bolshakov, A. Gelbukh. Computational linguistics: models, resources, applications. IPN – UNAM – Fondo de Cultura Económica, 2004.

[5] A. 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.

# ¿Qué gato tiene Juan?

*Juan usa un gato para reparar su carro.* ¿Qué gato?



Textos de entrenamiento:

| | |
|---|---|
| *Pedro usa un martillo para* | *el gato come ratones* |
| *Ana usa un desarmador para* | *el perro come la carne* |
| *el obrero usa una grúa para* | *el hámster come avena* |
| alguien usa éstos para algo | éstos, comen algo |

El gato de Juan ha de ser más parecido a un martillo, un desarmador o una grúa que a un perro o un hámster.

Diccionario monolingüe:

| | |
|---|---|
| *Martillo:* | *una herramienta que ...* |
| *Desarmador:* | *una herramienta que...* |
| *Grúa:* | *una herramienta que...* |
| *Gato* 1: | *un animal doméstico peludo.* |
| *Gato* 2: | *una herramienta que...* |

De las dos acepciones de gato, la segunda es la que más parece a martillo, desarmador o grúa. ¡Ya sabemos cuál gato!

Diccionario bilingüe: *Gato* 1) *cat*; 2) *jack.*

Ahora podemos traducir: *John uses a jack to repair his car*.

Imágenes: http://www.instrumentavto.ru/products_pictures/3.gif, http://overcoming.simfi.net/illustr/gazetta/kolobok.gif

Ilustración 1. Dónde se cultiva más la lingüística computacional

Ilustración 2. Los fundadores de la AMPLN: David Pinto Avendaño, Luis Villaseñor Pineda, César Antonio Aguilar, Azucena Montes, Luis Alberto Pineda Cortés, Andrés Soto, Manuel Montes y Gómez, Yulia Ledeneva, René Arnulfo García Hernández, Maya Carrillo, Alexander Gelbukh, Gerardo Sierra Martínez, Iván López-Arévalo, Concepción Pérez de Celis Herrero, Iván Vladimir Meza, Alberto Téllez Valero, Héctor Jiménez Salazar, Aurelio López López y Thamar Solorio quien tomó la fotografía.

Dr. Alexander Gelbukh es Maestro en Ciencias en las matemáticas y Doctor en las ciencias de la computación. Desde el 1997 es jefe del Laboratorio de Procesamiento de Lenguaje Natural del Centro de Investigación en Computación (CIC) del Instituto Politécnico Nacional (IPN). Es miembro de la Academia Mexicana de Ciencias, Investigador Nacional de México con nivel 2 y Secretario de la Mesa Directiva de la Sociedad Mexicana de Inteligencia Artificial (SMIA). Es autor, coautor o editor de más de 400 publicaciones y coautor de tres libros en las áreas del procesamiento de lenguaje natural e inteligencia artificial.