Procesamiento del lenguaje natural: estado de la investigación

Alexander Gelbukh
Centro de Investigación en Computación,
Instituto Politécnico Nacional, México

Resumen

El procesamiento del lenguaje natural es una rama de la ciencia que pertenece a la intersección de la lingüística aplicada y las ciencias de la computación, que estudia los métodos necesarios para que la computadora pueda ejecutar varias tareas relacionadas con el lenguaje humano, como el español, y requiere cierto grado de «entendimiento» de su contenido. Por otro lado esta ciencia desarrolla las herramientas que ayudan al lingüista en su trabajo cotidiano e incluso pueden llevar a descubrimientos lingüísticos nuevos. Este artículo presenta las aplicaciones principales del procesamiento automático de lenguaje natural y discuten los métodos que se emplean para la resolución de problemas. Al final del artículo se da una reseña bibliográfica para el lector interesado en obtener mayor información sobre el tema.

Introducción

En los últimos cinco-diez miles de años, la ocupación más importante de la raza humana ha estado enfocada a producir, mejorar y pasar-

les a las siguientes generaciones el tesoro más valioso que tenemos: el conocimiento. El conocimiento se almacena en la forma de lenguaje humano: libros, periódicos y en nuestros días la Internet. Como las computadoras son mucho más capaces que los seres humanos para manejar esta cantidad enorme de conocimiento, el estudio de las técnicas del procesamiento automático se ha vuelto una necesidad prioritaria y urgente.

La tarea final de la ciencia de la lingüística computacional —como la piedra filosofal para la alquímica— es la construcción de una máquina que pueda leer, escribir y conversar en nuestro lenguaje, ya sea español, inglés, chino, etcétera. Es un gran reto porque se trata de modelar en una computadora la actividad mental del ser humano. Aunque para bien o para mal cumplir con esta tarea no se ve como algo factible en el futuro cercano, al ir transcurriendo el camino hacia su tarea final, esta ciencia ya ha producido desarrollos muy útiles (al igual que hizo en su tiempo la alquimia) en tareas limitadas pero de gran importancia práctica.

A pesar de su nombre un poco confuso, la lingüística computacional no es una variante de la lingüística general; mientras ésta última estudia *qué* es el lenguaje, la primera estudia el *cómo* modelarlo computacionalmente con un fin práctico.[1] Se puede comparar esta relación con la que hay entre un ornitólogo y un constructor de aviones: el primero estudia los colores y las costumbres de las aves, el segundo tiene como tarea el *hacer* un ave —y no necesariamente que sea parecida en su estructura a las aves naturales, sino que vuele—.

Sin embargo existe una relación profunda entre las dos ciencias. La lingüística computacional, especialmente en la etapa «simbólica» de su desarrollo (véase más abajo), ha contribuido a la lingüística teórica[2] con los formalismos que ayudan a los lingüistas a expresar sus ideas con exactitud matemática y a comprobarlas experimentalmente sobre grandes corpus de textos. Con las herramientas computacionales los lingüistas pueden evitar el trabajo repetitivo y aburrido, y dejar que haga esto la máquina. Los fenómenos descubiertos durante el

---

1 En este sentido se puede clasificar como una rama de la lingüística aplicada.

2 O ha inspirado su creación.

desarrollo de los sistemas computacionales aplicados han sido una inspiración para los descubrimientos lingüísticos. Por otro lado, la lingüística teórica ha ayudado profundamente, con sus métodos y sus conceptos a la lingüística computacional; las dos ciencias se complementan sin sustituirse.

Desde el punto de vista técnico la tarea del procesamiento automático del lenguaje es (viéndolo por parte del análisis) la definición de la estructura en el texto dado. El mayor problema en este camino es la ambigüedad: no es tan difícil encontrar una estructura en un segmento del texto en la que se tenga que elegir entre más de una estructura posible, al existir diferentes interpretaciones de ese segmento. Por ejemplo, la palabra «sobre» puede ser un verbo, un sustantivo o una preposición («todo lo que sobre se dejará en un sobre sobre la mesa»); la frase «veo al gato con un telescopio» se puede interpretar como «veo al gato, y para eso uso un telescopio» o «veo al gato, el cual tiene un telescopio». Tales ambigüedades se presentan en grandes cantidades en casi cualquier oración o palabra, aunque el lector humano raras veces las nota. Para resolver tales ambigüedades es necesario o bien analizar un contexto más amplio del segmento de texto dado, o bien emplear el conocimiento externo al texto dado: por ejemplo, la ambigüedad con la palabra «sobre» se puede en muchos casos resolver analizando su contexto: «El documento fue entregado en un sobre cerrado».

En este artículo presentamos brevemente las aplicaciones principales del procesamiento automático de texto, discutimos las tareas específicas que técnicamente aparecen en el proceso de análisis, describimos los tipos de métodos principales que se ocupan para resolver los problemas y adjuntamos una bibliografía para el lector interesado en obtener mayor información sobre el tema.

Aplicaciones

Mencionaremos aquí sólo tres tipos de las aplicaciones principales del procesamiento del lenguaje natural.

Búsqueda de información. La forma más tradicional de búsqueda de información es la *recuperación de información.* Se trata del es-

cenario en el cual el usuario especifica los criterios que deben satisfacer los documentos de interés —usualmente unas palabras que éstos deben contener— y el sistema debe proporcionarle estos documentos. El nombre de la aplicación, «recuperación de información», refleja su forma más básica e históricamente primera: el proporcionar, o «recuperar», el conjunto de los documentos que satisfacen el criterio. Sin embargo las técnicas modernas de recuperación de información no implican selección sino ordenamiento de toda la colección de documentos disponible siguiendo el mayor o menor grado de satisfacción del criterio, así que los documentos que mejor satisfacen el criterio aparecen al inicio de la lista.

Dos consideraciones son importantes aquí. Primero, en este escenario el criterio no simplemente se satisface o no, sino que puede satisfacerse en mayor o menor grado, por ejemplo, el documento puede contener mayor o menor número de palabras relacionadas con la petición del usuario, aquí las técnicas del procesamiento del lenguaje natural ayudarían a decidir qué palabras y en qué contextos deberían relacionarse con la petición. Segundo, y más importante, la tarea se entiende no sólo como la comparación entre la petición del usuario con los textos de los documentos, sino como satisfacción de la necesidad informática del usuario. Esto involucra el razonamiento lógico o estadístico sobre la necesidad del usuario y la mejor manera de satisfacer ésta.

Al satisfacer la necesidad del usuario las técnicas modernas van más allá de simplemente proporcionarle una lista ordenada de los documentos. Por el tipo de la necesidad y el uso de la información que se quiere obtener, se distinguen las siguientes corrientes modernas en la búsqueda de la información.

En la *respuesta a preguntas,* la petición del usuario se entiende como una pregunta («¿dónde se organizan los juegos Olímpicos en 2008?») y la salida del sistema se da no en la forma de una lista ordenada de documentos sino en la forma de una simple respuesta a tal pregunta («en China»). Esto le ahorra tiempo al usuario ya que en lugar de tener que analizar los documentos, él recibe la respuesta directa. En este tipo de técnicas el procesamiento del lenguaje natural es necesario para «entender» tanto la pregunta del usuario como la información

presentada en los documentos de la colección en la cual se busca la respuesta. La corriente que activamente se desarrolla en la actualidad es la respuesta a preguntas multilingües: el lenguaje de la pregunta (por ejemplo, español) puede ser diferente del lenguaje de los documentos analizados (por ejemplo, inglés), o incluso éstos pueden estar en diferentes lenguajes; la situación es importante para la Unión Europea y varios otros países.

En la *generación de resúmenes*, el usuario está interesado en obtener una idea general sobre un documento antes de leerlo, o para decidir si lo necesita leer. El sistema entonces le proporciona una representación resumida del documento. Esto es especialmente importante cuando se aplica a una colección de muchos documentos, posiblemente millones, los cuales el usuario simplemente no podría leer. Aún más difícil es la tarea cuando se trata de documentos escritos en diferentes lenguajes, o cuando el lenguaje del resumen requerido no coincide con el lenguaje de los documentos.

En la *minería de texto,* también se busca una especie de representación resumida de una gran cantidad de textos, pero en este caso tal «resumen» no refleja lo que los autores querían expresar en sus textos sino la metainformación sobre estos textos: la información de las tendencias, opiniones prevalecientes, porcentajes de diferentes opiniones, etcétera. Gracias a muchas aplicaciones prácticas en la política, economía, comercio, sociología y otras áreas de la vida social, esta rama del procesamiento del lenguaje natural ha recibido recientemente una gran atención por parte de los investigadores y las empresas.

Traducción automática. Con la globalización, la abundancia de la información textual en la Internet, la intensificación de los contactos internacionales y la disponibilidad de Internet por cada vez un mayor número de personas en los países menos desarrollados convirtieron a la traducción automática en una aplicación prioritaria. Debido a que la traducción requiere un procesamiento detallado y profundo del significado del texto, ésta fue una de las motivaciones principales que desarrollaron la ciencia de la lingüística computacional desde la época de las primeras computadoras electrónicas.

El avance de la traducción automática en los últimos años ha sido impresionante: los traductores automáticos pasaron de una herra-

mienta auxiliar o más bien experimental a los sistemas útiles para el uso masivo cotidiano de millones de usuarios no expertos. Uno de los mejores traductores automáticos disponibles en la actualidad es el de *Google*©.

Interfaces en lenguaje natural. Esto será probablemente la tarea del futuro, y hasta ahora ha logrado un avance limitado. La idea es que en el futuro cuando las computadoras empiecen a actuar en el mundo real (robots inteligentes) y ayudar en sus quehaceres cotidianos a la gente no experta, será indispensable que puedan hablar con los usuarios en lenguaje natural. Sin embargo tal cosa requiere un alto grado de entendimiento del mensaje, y además involucra áreas de la lingüística computacional que todavía están en pañales: el manejo del diálogo y de ciencia liguistíca de la pragmática, entre otras.

Es por eso que el manejo del diálogo atrae la atención de muchos investigadores y es un área de investigación activa. No podríamos decir lo mismo de la pragmática computacional, todavía se conoce muy poco sobre el tratamiento computacional de la pragmática.

Tareas

Para lograr el objetivo requerido, los sistemas de procesamiento automático de texto dividen internamente el trabajo de una u otra manera en bloques cuya información de entrada y de salida está bien definida, de tal manera que el sistema completo se puede armar de estos bloques intercambiables. En la mayoría de los casos se trata de la ejecución consecutiva de los bloques, donde la salida del bloque anterior sirve como entrada al siguiente. En muchos casos, cada bloque toma en la entrada una secuencia de elementos más simples y agrupa algunos de ellos produciendo una secuencia de elementos compuestos más «grandes».

Tales bloques permiten el estudio independiente y la especialización de los grupos de investigación, los investigadores, los libros y los recursos computacionales y los léxicos. En muchas tareas se han establecido incluso prestigiosas competencias internacionales donde se definen los mejores sistemas para la tarea en cuestión. A las correspondientes ramas de la lingüística computacional se les refiere como

tareas (*tasks* en inglés). No todas las tareas se emplean en una aplicación dada. Usualmente se emplean las tareas hasta un cierto nivel necesario. Podemos mencionar los siguientes ejemplos de tareas:

Preprocesamiento. Es un nombre paraguas para varias tareas que se consideran relativamente simples —más simples que las tareas que le siguen— y que se ejecutan antes de las tareas más complicadas. Diferentes autores pueden considerar algunas de las subtareas aquí mencionadas como tareas independientes.

Limpieza del texto es la localización y eliminación o el tratamiento adecuado de los segmentos del archivo de texto que no son propiamente texto: imágenes, marcaje del formato, fórmulas, letras mayúsculas y minúsculas, etcétera; es decir, todo aquello que ayuda a convertir el archivo dado en un texto bien formado en el sentido lingüístico.

Determinación de elementos textuales (*tokenizing* en inglés) implica convertir el texto de una secuencia de letras a una secuencia de cadenas, aproximadamente correspondientes a palabras. Es importante recalcar que esta tarea en apariencia tan simple está en realidad interconectada con decisiones que se toman en otros niveles. Por ejemplo, en la secuencia de caracteres «etc.», ¿es el punto parte de la palabra o es un elemento textual aparte que indica la finalización de la oración? El programa que efectúa esta tarea se llama en inglés *tokenizer*.

Reconocimiento de entidades nombradas (NER, *named entity recognition* en inglés) es la etapa de aglutinar varias palabras en una sola que es el nombre de un concepto u objeto; por ejemplo. «Estados Unidos Mexicanos» se tratará como una sola palabra que refiere al país correspondiente y es sinónima de «México». Se aplican las mismas aclaraciones sobre la no trivialidad de esta tarea. El descubrimiento de tales entidades nombradas y la toma de decisiones sobre la aglutinación, o no, de los elementos textuales en un contexto dado en una entidad nombrada, han sido áreas de activa investigación en los últimos años.

Separación de las oraciones (*sentence splitting* en inglés) es la toma de decisión acerca de dónde están las fronteras entre las oraciones. El ejemplo arriba mencionado de la secuencia «etc.» ilustra un ca-

so no trivial de esta tarea. Un programa que efectúe esta tarea se llama en inglés *sentence splitter.*

Eliminación de las palabras basura (*stopwords* en inglés). Para ciertas aplicaciones, las palabras sin contenido semántico, tales como las palabras funcionales o muy frecuentes, no aportan casi nada al desempeño del sistema y pueden perjudicarlo. La lista de las palabras que deben eliminarse puede variar de una aplicación a otra y en ocasiones está sujeta a decisiones inteligentes automáticas.

Análisis morfológico. Esta tarea de índole propiamente lingüística involucra la determinación de la primera forma de la palabra y sus características gramaticales y morfosintácticas. Se puede efectuar sin contexto, es decir, sobre una palabra aislada, y en este caso proporcionará un resultado ambiguo, por ejemplo: «como», «nada», «habla», «hablamos». También se puede efectuar en el contexto, lo que usualmente permite desambiguar la palabra para dar una sola respuesta: «considerarlo como tal», «Juan nada bien», «él habla poco», «ayer hablamos de eso». El programa que efectúa la tarea completa sin contexto —dando usualmente varias variantes de la respuesta— se llama un analizador morfológico, y si lo hace en contexto —produciendo sólo una variante de la respuesta— se llama etiquetador (*tagger).* El programa que sólo determina la primera forma de la palabra dada («comió» «comer»), se llama en inglés *lemmatizer.* Para ciertas aplicaciones, sobre todo en la recuperación de información, es suficiente determinar sólo la base de la palabra sin generar la forma completa («comió» «com-»).

Análisis sintáctico. Se trata de agrupar las palabras dentro de una oración, según su relación sintáctica y descubrir así su estructura interna. Existe un número considerable de teorías y metodologías para efectuar esta tarea, las cuales difieren substancialmente en la definición de ella y la definición de la relación sintáctica. Las corrientes más comunes y más distintas entre sí son la corriente de análisis sintáctico de constituyentes, y la de dependencias.

En la primera las palabras se agrupan y se forma un árbol sintáctico cuyos vértices son grupos de palabras relacionadas, y las aristas son las relaciones de anidamiento entre estos conjuntos. En la segunda aproximación las palabras se asignan una como «ayudante»

de la otra, y esta relación forma un árbol cuyos vértices son palabras individuales. Por lo general el análisis de constituyentes tiene un mejor sustento matemático y es más fácil de efectuar, mientras que el análisis de dependencias proporciona información más rica y mejor organizada para el subsecuente análisis semántico. Los programas que efectúan el análisis sintáctico se llaman en inglés *parsers,* aunque este término más comúnmente se aplica al análisis de constituyentes.

Análisis semántico. Esta tarea concierne al tratamiento del significado del texto más que a su estructura, y puede involucrar varios aspectos del problema.

La *desambiguación de los sentidos de las palabras* (WSD: *word sense disambiguation* en inglés) es la tarea de elegir una sola acepción de la palabra, del repertorio dado por un diccionario, dependiendo del contexto dado. Por ejemplo las diferentes acepciones que se le deben asignar a la palabra banco en los contextos «Juan tiene cuenta en el banco», «la arena del banco del río Nilo» y «el banco de peces». La tarea es muy importante hasta en las aplicaciones básicas que no requieren del análisis semántico completo, tales como la recuperación de información, y recientemente recibe mucha atención. Se organiza una competencia internacional sobre la tarea, SemEval.[3] En el sitio de SemEval se pueden obtener las colecciones y corpus necesarios para la investigación en esta tarea.

La *resolución de anáfora* es un caso extremo de la desambiguación del sentido, cuando la palabra misma —un pronombre— prácticamente no da ninguna pauta para decidir qué se entiende por esta palabra, sino toda la información viene de su contexto. Por ejemplo: «Juan tomó la torta de la mesa y la comió», ¿qué comió Juan? En ciertos contextos, a veces ni siquiera el pronombre aparece explícitamente en el texto: «Juan compró una casa. La cocina es grande», ¿cuál cocina? [4].

La *detección de correferencia* es una aplicación de tarea relacionada con la resolución de anáfora: es la tarea de detectar las palabras que refieren a la misma entidad o acción en el mundo descrito en el texto, por ejemplo: «María vió a Juan. El hombre se asustó. «Sr. Sánchez, soy yo», le dijo.»

---

3 www.SENSEVAL.org

La *detección de la implicación lógica* (*textual entailment* en inglés), es una tarea que ha recibido gran atención en los últimos años, y este interés sigue creciendo rápidamente. Se considera una tarea unificadora de muchas áreas de la lingüística computacional; es decir, una tarea que implica en cierta forma a muchas otras tareas y aplicaciones del procesamiento del lenguaje natural, tales como la desambiguación de los sentidos de las palabras, la resolución de anáfora, la recuperación de información, la respuesta a preguntas. Esta tarea trata de responder «sí» o «no» a una simple pregunta: basándose en dos textos, ¿el primero implica al segundo? Por ejemplo, «Juan vive en Paris», ¿implica «Juan vive en Francia»?[4]; «Juan se murió» ¿implica «Juan vivió»? Es fácil ver que la tarea trivialmente implica varios tipos de desambiguación: «Juan tomó la torta de la mesa y la comió», ¿implica «Juan comió la torta» o «Juan comió la mesa»?; «Juan trabaja en el banco» ¿implica «Juan trabaja en una organización financiera» o «Juan trabaja en una orilla del río»?

Finalmente, el *análisis semántico* propiamente dicho busca la descripción explícita de las relaciones lógicas entre las palabras, el significado y la interpretación de tales palabras. Una parte de esta tarea es la construcción de una red de las relaciones lógicas entre las entidades, las acciones y las propiedades mencionadas en el texto. Las palabras correferentes, corresponden al mismo vértice de tal red, el cual hereda las relaciones en las cuales está involucrada cada una de estas palabras en el texto. A pesar de un esfuerzo significativo destinado a esta tarea, y relativamente poco lo que se ha logrado. Existe un gran número de teorías al respecto, lo que indica que no existe ninguna lo suficientemente buena.

Análisis pragmático. Mientras que el análisis semántico todavía trata de estructurar lo que directamente dice el texto, el análisis pragmático proporciona metainformación sobre el texto: para qué se dice lo que se dice, cómo se organiza, qué estrategias emplea el autor para alcanzar el efecto deseado.

4 Estrictamente, la respuesta es «no», porque el topónimo puede referir a una de veinte ciudades en el mundo con este nombre, sin mencionar interpretaciones más exóticas como un virus llamado Juan que vive en Paris Hilton. Sin embargo, en la práctica usualmente se espera una interpretación razonablemente probable.

El *análisis de texto* es una evaluación del texto como una entidad completa, no dividida en oraciones. En esta fase se pretende entender el efecto que quería lograr el autor al comunicarle este texto al lector.

El *análisis de discurso* determina el papel lógico de cada parte del texto —oración, frase u otro fragmento— en la presentación completa de los argumentos del autor. Por ejemplo, en el texto «El libro es muy interesante. Juan lo leyó en una noche» la primera oración es la explicación para la segunda: lo leyó *porque* es interesante.

El *análisis y planeación de diálogo* es necesario cuando no se trata de un texto (monólogo) sino de intercambio de intervenciones de varias personas. Esta situación involucra los elementos que un texto normal no presenta, por ejemplo, la conservación de la información de un turno a otro, duración de los turnos y las estrategias para ceder o solicitar su turno, la estructura de los estímulos y las respuestas entre los participantes, etcétera.

Finalmente, el *análisis pragmático* propiamente dicho es la determinación de las intenciones del autor o hablante: para qué se dice lo que de dice, qué efecto espera lograr el autor con su texto. Por ejemplo, pragmáticamente con la frase «¿me puedes pasar la sal?» se pretende causar al escuchante pasar la sal, mientras que semánticamente se trata de una solicitud de información sobre las habilidades físicas del escuchante de hacerlo.

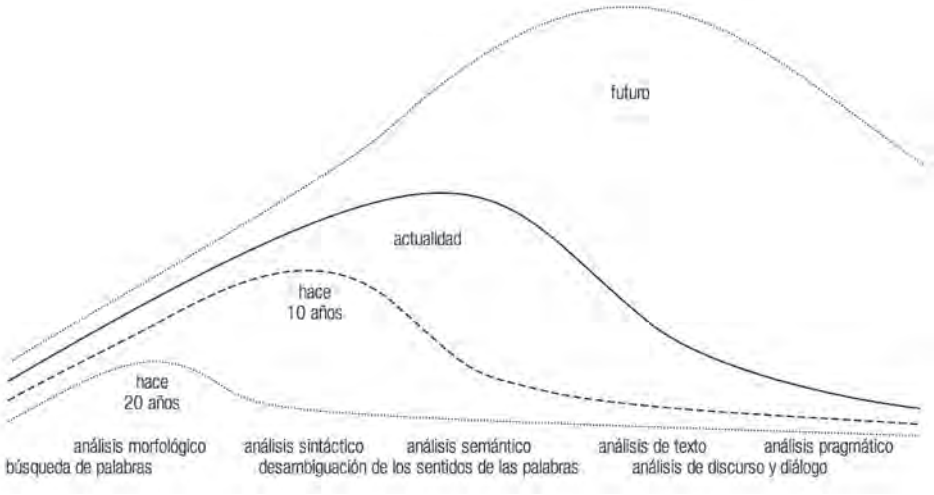
Las tareas enumeradas en esta sección pueden parecer simples, y cualquier niño las hace sin problemas. Sin embargo, cuando se intenta que las haga un programa de manera totalmente automática y autónoma, se presentan dificultades técnicas tan considerables que a pesar de todo el esfuerzo dedicado a la investigación correspondiente, el avance en muchas de estas tareas es todavía cuestión del futuro cercano o no tan cercano.

Históricamente en las etapas iniciales del desarrollo de la lingüística computacional, la mayor atención la han recibido las tareas más simples.[5] Con el tiempo estas tareas pasan a ser consideradas como más o menos resueltas; mientras tanto la mayoría de los investigado-

5 La tesis de doctorado del autor de este artículo, iniciada hace 20 años, fue en el área del análisis morfológico.

res se enfocan en las tareas más complejas que reflejan un mayor nivel de profundidad de entendimiento del lenguaje; es decir, la estructuración de mayores segmentos del texto o la consideración de un contexto más amplio, véase la Ilustración 1.

Ilustración 1. La evolución histórica aproximada de la inversión del esfuerzo en las tareas de diferentes niveles de profundidad. Obsérvese la evolución del área del mayor esfuerzo invertido, y la mayor actividad de la investigación al ir desde tareas más simples hacia las más complejas.



Métodos simbólicos y estadísticos

La lingüística computacional nació hace medio siglo como una rama de la lingüística que pretendía dar la descripción de los fenómenos lingüísticos o las recetas de análisis lingüístico del texto con tal claridad que incluso una máquina pudiera aplicar dicha información para realizar las tareas prácticas. Cuando en la escuela se le enseña al niño cómo hacer un análisis formal de la oración, la maestra da ejemplos de tal análisis o explicaciones poco específicas, poco detalladas y poco claras, y la interpretación de tal explicación se le deja a la intuición lingüística natural del niño. La práctica de la presentación del material

en los libros científicos en las humanidades tradicionalmente es similar: se le deja mucho a la intuición y a la interpretación del lector.

Pero para la computadora tal estilo de descripción no es aceptable —ella necesita reglas claras, con todas las posibles opciones explícitamente mencionadas, y con la definición de los conceptos usados en tales reglas. De hecho, esto es sano también para la misma ciencia: lo que es claro para la computadora también será claro tanto para el mismo autor como para los lectores. Además, varios fenómenos nuevos de gran importancia fueron descubiertos sólo tras una introspección en la búsqueda de la claridad completa: los fenómenos de los que los científicos no se daban cuenta porque nunca se hicieron preguntas tales como «¿cómo decido cuál palabra elegir?» o «¿qué específicamente significa el concepto que parece claro cuando veo los ejemplos?»

La lingüística computacional nació así como una ciencia de las reglas claras formuladas por los expertos humanos.[6] Al enfoque basado en las condiciones lógicas que o bien se cumplen o bien no (no se pueden cumplir parcialmente), y en los conceptos discretos que o bien son iguales o bien diferentes (no pueden ser «parecidos»), se les conoce como el enfoque *simbólico.* Con este enfoque resultó posible efectuar el análisis morfológico, sintáctico, semántico (hasta cierto grado), desarrollar los sistemas de recuperación de información e incluso los sistemas de traducción automática de alta calidad.

Sin embargo este enfoque tiene algunas deficiencias. Primero, el costo del desarrollo de un sistema de reglas lo suficientemente complejo se vuelve prohibitivamente alto. Segundo, el número de variantes que generan tales reglas en el análisis de un fenómeno ambiguo rápidamente se vuelve inmanejable.[7] El problema es que con las reglas de tipo lógico es difícil distinguir entre las variantes probables y las menos probables: un humano difícilmente puede razonar en términos de probabilidad, ni tampoco puede estimar correctamente las probabilidades de diferentes fenómenos. Los sistemas basados en el

---

6 Fue en los tiempos del florecimiento de los sistemas expertos en la inteligencia artificial, los cuales también constituyen una formulación clara del conocimiento del experto en el área.

7 Véase más arriba el ejemplo de un virus llamado Juan que vive en Paris.

enfoque simbólico resultaron ser demasiado rígidos y poco robustos: el sistema o bien da millones de variantes o bien —con más restricciones— ninguno.

Con la disponibilidad de enormes cantidades de textos en Internet, hace aproximadamente una década,[8] se descubrió la posibilidad de extraer la información sobre las palabras y los fenómenos lingüísticos a partir de los textos. Esto resultó una mina de oro para la investigación en la lingüística computacional y las técnicas de procesamiento del lenguaje natural avanzaron mucho. Los métodos estadísticos no necesitan el costoso esfuerzo por parte de los expertos líderes para construir los recursos léxicos y las gramáticas.

Más aún, los recursos construidos tienden a ser no cualitativos, como en el caso del enfoque simbólico, sino cuantitativos. No sólo se sabe que existe una palabra sino cuántas veces aparece; no sólo se sabe que existe una regla gramatical sino cuántas veces se usa; no sólo se sabe que dos palabras son sinónimas sino que se puede medir cuantitativamente el grado de su similitud. Esta información habilita el razonamiento estadístico no disponible para los recursos con la información cualitativa.

Como un ejemplo de tal razonamiento, consideremos el siguiente razonamiento falaz: P ⇒ Q, Q, entonces P.[9] Se convierte en razonamiento válido cuando se trata de probabilidades: P es posible y Q es poco probable; dado P es muy probable que Q, Q, entonces muy probable que P,[10] lo que se sustenta en la regla de Bayes:

$$\mathrm{P}(P \mid Q) = \mathrm{P}(Q \mid P)\frac{\mathrm{P}(P)}{\mathrm{P}(Q)},$$

donde P(P) es la probabilidad de P y P($P \mid Q$) es la probabilidad de P dado Q.

8 Obviamente había trabajos correspondientes desde hace dos o tres décadas, pero no eran la corriente prevalente.

9 Si Juan mató a Pedro, entraría en la casa de Pedro. Lo vieron entrar en su casa. Entonces, él lo mató.

10 Juan tenía enemistad con Pedro y nunca antes venía a su casa. Aquella noche lo vieron entrar en la casa. Entonces, es muy probable que él lo haya matado.

Este tipo de razonamiento les permite a los enfoques estadísticos resolver las ambigüedades que los enfoques simbólicos no pueden resolver. Por ejemplo, sí se puede afirmar con alto grado de certeza que si Juan está en París, entonces está en Francia, a menos que el contexto presente evidencias de que se trata de otro lugar llamado París[7] —lo que es P(*Q*). Hay que aclarar que el problema de los enfoques simbólicos no es que no puedan aplicar el razonamiento estadístico, sino que por lo general no disponen de la información necesaria para ello: los expertos humanos no son capaces de proporcionar tal información con un grado razonable de precisión.

Es por eso que los métodos estadísticos constituyen el tema de la gran mayoría de los trabajos de investigación en nuestros días, mientras que los métodos simbólicos casi han caído en el olvido durante la última década. Así la lingüística computacional contemporánea ha dejado de ser una rama de la lingüística, por lo menos de la lingüística en su forma tradicional, y se ha convertido en un alto grado en una rama del aprendizaje automático, un área de la ciencia de la Inteligencia Artificial.[11]

En la ciencia del aprendizaje automático se distinguen dos tipos de aprendizaje: el aprendizaje supervisado —basado en ejemplos— y el aprendizaje no supervisado. El primer tipo de aprendizaje requiere de un corpus de textos donde la tarea ya está resuelta: por ejemplo, para la tarea del etiquetado de las categorías gramaticales (*tagging*) la categoría gramatical de cada palabra ya esté marcada manualmente. El sistema entonces infiere las reglas usadas (quizá inconscientemente) por el individuo que hizo tal etiquetado, y luego puede continuar con la tarea sobre los textos nuevos, que no tienen el etiquetado.

La ventaja es que el anotador no tiene por qué ser experto en lingüística ni entender las razones subyacentes de sus acciones, sino que puede ser cualquier hablante nativo educado del lenguaje dado, al menos cuando se trata de tareas tan simples como el etiquetado de las categorías gramaticales. Las desventajas principales son dos: primero, todavía se necesita una cantidad enorme de trabajo manual, aunque

11 véase www.MICAI.org, www.SMIA.org.mx

no de un experto; segundo, lo que se descubre en este proceso no es la información sobre el lenguaje sino la información sobre las opiniones subjetivas del anotador; se estudia la cabeza del anotador y no el lenguaje.

Estos problemas son superados en el otro tipo de aprendizaje —el aprendizaje no supervisado— que son las técnicas que no requieren preparación alguna de los textos sino que toman como entrada los textos planos tal cual se encuentran en, digamos, Internet. El resultado del aprendizaje no supervisado se parece a lo que hemos visto en las películas de ciencia ficción, cuando un marciano sale de su plato volador, graba con un dispositivo las palabras que le dirigen los terrícolas, y en un rato el dispositivo empieza a traducirlas al marciano y al revés. Sorprendentemente, las técnicas del aprendizaje estadístico, sobre todo el no supervisado, hacen tal escena cada vez más parecida a ciencia que a ficción [9].[12]

Sin embargo, la marcha triunfal de los métodos estadísticos no implica que no tengan sus deficiencias. Primero, con estos métodos no se aprovecha el conocimiento de los expertos en lingüística y la estructuración tradicional del objeto de nuestros estudios. Segundo, los métodos estadísticos son menos eficientes en el descubrimiento de estructuras muy complejas; como son, por ejemplo, las estructuras semánticas. Se puede esperar que esta mina de oro podría empobrecer pronto, lo que haría atractivo para los investigadores el hecho de combinar los métodos simbólicos, basados en la experiencia e ingenio humanos, con los métodos estadísticos, que pueden proveerles las características probabilísticas y cuantitativas.

Una reseña bibliográfica

Mayor información sobre los temas presentados en este capítulo se puede encontrar en varios libros, revistas, y memorias de congresos.

A los lectores que prefieren leer en español, como una ilustración de la investigación en el campo se les puede recomendar los libros

12 véase también www.reviews.com/review/review_review.cfm?review_id=129833

[3, 5], disponibles para su descarga desde las direcciones indicadas. El libro [3] describe una aproximación prevalentemente simbólica del análisis sintáctico del español, con intentos de combinarlo con aproximaciones estadísticas. El libro [5] incluye varios trabajos de índole prevalentemente estadística, de análisis computacional del corpus y de extracción automática de los recursos léxicos de los textos disponibles. Está en preparación la traducción al español del libro [3], la cual aparecerá en la dirección indicada cuando esté terminado; se puede obtener un borrador directamente de los autores. Artículos de investigación de los mismos autores se pueden encontrar en el servidor del Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC, IPN.[13]

Otra fuente de información son las memorias y las ponencias de varios congresos nacionales organizados en México sobre el tema. Uno de éstos es el Taller Nacional de Tecnologías del Lenguaje Humano, organizado en distintas ciudades del país por el Laboratorio de Tecnologías del Lenguaje del INAOE, Puebla. Otro es el Coloquio de Lingüística Computacional de la UNAM, organizado por el grupo de Ingeniería Lingüística. En ambos eventos se puede conocer en persona a los expertos y los estudiantes nacionales que trabajan en el tema, y familiarizarse con sus líneas de trabajo y sus proyectos actuales. En México también está en proceso de consolidación la Asociación Mexicana de Procesamiento del Lenguaje Natural, AMPLN.[14]

La revista *Procesamiento del Lenguaje Natural* editada por la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN,[15] publica artículos originales de investigación tanto en español como en inglés, orientados hacia el lector preparado o experto. Sin embargo, un lector que recién se inicia en el tema puede encontrar información interesante sobre la actualidad de investigación y sobre los métodos empleados, así como ubicar a las personas y los grupos españoles y latinoamericanos relevantes. Los artículos de la revista están disponibles desde el servidor de la Asociación.

---

13 NLP.CIC.IPN.mx

14 www.AMPLN.org

15 www.SEPLN.org

Los lectores interesados en leer en inglés tienen mucho mayores posibilidades para obtener la información. Una introducción relativamente concisa al tema es el [2]. Al lector avanzado se le puede recomendar un libro de texto muy detallado el [6], el cual probablemente no sea necesario leer completo sino sólo las partes relevantes para el trabajo propio del lector. En cuanto a la aproximación puramente estadística al procesamiento del lenguaje, el libro clásico y también muy detallado es el [8]. A los lectores interesados específicamente en la recuperación de información, se les puede recomendar el clásico libro [1] y el más moderno [7], que presentan tanto el material básico como, en sus últimos capítulos, el de vanguardia.

Existen varias revistas de alta calidad dedicadas al procesamiento del lenguaje natural, entre las cuales destacan *Computational Linguistics, Research in Language and Computation, Language Resources and Evaluation, ACM Transactions on Speech and Language Processing, Natural Language Engineering, Journal of Quantitative Linguistics,* por mencionar algunas. Estas revistas publican artículos de investigación originales, orientados hacia los expertos.

Un panorama más amplio se puede ver en los mejores congresos, los cuales discuten los temas de investigación de frontera. La estructura de las memorias de un buen congreso da una idea clara de las direcciones de investigación que en la actualidad son las más activas y son una muestra representativa sobre las preguntas de investigación y los métodos que se aplican. Se le puede recomendar al lector que revise las memorias del congreso internacional CICLing,[16] organizado en México —y en ocasiones en otros países del mundo— por el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC, IPN[13]. Desde el sitio Internet del congreso están disponibles las tablas de contenido de sus memorias de los 10 años de su historia, así como los resúmenes de los artículos. La lectura de los artículos de un congreso de esta naturaleza es probablemente la mejor manera para familiarizarse con el estado actual del área.

Otros excelentes congresos internacionales en el área son los organizados por la *Association for Computational Linguistics*

16 www.CICLing.org

(ACL):[17]ACL, NAACL, EACL, así como COLING,[18] RANLP, TSD[19] y NLDB.[20] Las memorias de estos congresos también están disponibles, ya sea en texto completo o sólo los resúmenes.

Finalmente, los postgrados en México y en el extranjero ofrecen los programas tanto del nivel de Licenciatura (UNAM[21]) como del nivel Maestría y Doctorado (CIC-IPN,[13] INAOE,[22] UNAM[21,23]).

Conclusiones

En este artículo hemos presentado las aplicaciones principales que en la actualidad trabaja el área del procesamiento automático de lenguaje natural, y hemos discutido también los métodos e ideas principales que se emplean para resolver los problemas en esta ciencia.

El procesamiento del lenguaje natural, llamado también la lingüística computacional, es una rama de la ciencia que se encuentra en la intersección entre la lingüística aplicada y la ciencia de la computación. La lingüística computacional y la lingüística teórica son dos ciencias que mutualmente se complementan y cada una aporta valor para el desarrollo de la otra.

En la etapa pasada la lingüística computacional se ocupaba principalmente de codificar claramente el conocimiento lingüístico, y en este sentido fue, como lo indica su nombre, una rama de la lingüística. En la etapa contemporánea se ha convertido en una rama de la ciencia de la inteligencia artificial; a saber, el aprendizaje automático. Un reto muy interesante es el aprendizaje no supervisado (de los textos disponibles no marcados manualmente) del conocimiento lingüístico y de los datos (diccionarios y gramáticas) necesarios para desarrollar los sistemas prácticos.

---

17 www.aclweb.org
18 nlp.shef.ac.uk/iccl
19 www.tsdconference.org
20 www.nldb.org
21 www.iingen.unam.mx
22 ccc.inaoep.mx/labtl
23 turing.iimas.unam.mx

De acuerdo con esto existen dos corrientes en la lingüística computacional: la corriente simbólica y la estadística. La primera involucra el conocimiento humano y opera con los datos cualitativos. En esta corriente, las ambigüedades del análisis del texto se resuelven a través del análisis de un contexto más amplio y en los niveles más profundos del análisis: por ejemplo, la ambigüedad de la categoría gramatical de la palabra se resuelve durante el análisis sintáctico de la oración completa; la ambigüedad sintáctica de la oración se resuelve en la fase del análisis semántico del texto completo.

El enfoque estadístico tiende apoyarse en el aprendizaje automático de los datos necesarios y a operar con los datos cuantitativos utilizando los cálculos probabilísticos para resolver las ambigüedades. Este enfoque es el prevaleciente en la etapa contemporánea y ha producido resultados excelentes. Sin embargo tiene sus limitaciones, aunque se espera que en el futuro los dos enfoques habrán de combinarse.

Al lector interesado en obtener más información se le invita a consultar los libros de texto mencionados arriba, o las memorias de los mejores congresos, tales como el ACL o el CICLing, o bien integrarse en uno de los grupos de investigación o programas de posgrado existentes en el país.

Referencias

R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.

I. A. Bolshakov, A. Gelbukh. *Computational Linguistics: Models, Resources, Applications.* IPN-UNAM-FCE, 2004, 186 pp.; www.gelbukh.com/clbook.

S. N. Galicia Haro, A. Gelbukh. *Investigaciones en análisis sintáctico para el español.* IPN, 2007, 324 pp.; www.gelbukh.com/libro-investigaciones.

A. Gelbukh, G. Sidorov. *On Indirect Anaphora Resolution.* Proc. PACLING-99, Canada, 1999, pp. 181–190; www.gelbukh.com/CV/Publications/1999/PACLING-1999-Anaphora.htm.

A. Gelbukh, G. Sidorov. *Procesamiento automático del español con enfoque en recursos léxicos grandes.* IPN, 2006, 240 pp.; www.gelbukh.com/libro-procesamiento.

D. Jurafsky, J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Published by Pearson Prentice Hall, 2008, 1024 pp.

C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

C. Manning, H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press. Cambridge, MA, 1999.

D. W. Oard. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)* 2, 2, 2003, 79-84.