Procesamiento del lenguaje natural, un reto de la inteligencia artificial

Natural language processing
A challenge for artificial intelligence

Chaves Torres, Anívar[1]
Universidad Nacional Abierta y a Distancia UNAD
Escuela de Ciencias Básicas, Tecnología e Ingeniería
anivarchaves@yahoo.com

Zuleta Medina, Alejandra[2]
Institución Universitaria CESMAG
alejazul07@gmail.com

Recibido: 30 de noviembre de 2012
Aceptado: 05 de octubre de 2013

"El lenguaje es el bien más precioso
y a la vez el más peligroso
que se ha dado al hombre."
Hölderlin

Resumen **—** Este artículo presenta una comparación entre el lenguaje natural y los artificiales, resaltando que el lenguaje humano es complejo, multiforme y rico en expresiones, pero a la vez, ambiguo, requiriendo interpretación de acuerdo al contexto y a la intensión del hablante; mientras que los lenguajes artificiales, se diseñan con una finalidad concreta, son restringidos tanto en la sintaxis como en la semántica, razón por la cual son más precisos, con menos espacio para la libre interpretación y libres del contexto. Se muestra la importancia de la investigación sobre el procesamiento automático del lenguaje natural, se referencian algunos de los principales avances en este campo y las áreas donde se requiere procesar el lenguaje natural para mejorar los sistemas informáticos.

Palabras clave **—** Lenguaje natural, lenguaje artificial, procesamiento automática del lenguaje.

Abstract **—** This paper presents a comparison between the natural language and the artificial languages, highlighting that human language is complex, multifaceted and rich in expressions, yet, is ambiguous and requires interpretation according to the context and intent of the speaker, while that artificial languages are designed for a specific purpose, are limited in their syntax and semantics, thus are more accurate, with less space for free interpretation and context free.
It shows the importance of research on the automatic processing of natural language, is referenced several major advances in this field and the areas which require natural language processing to improve computer systems.

Keywords **—** Natural language, artificial language, automatic language processing.

I. INTRODUCCIÓN

La comunicación es una facultad principalmente de los seres humanos, que ha contribuido enormemente en la organización y desarrollo de la sociedad. Esta se lleva a cabo a través del lenguaje en sus diferentes formas. Últimamente, con ocasión del auge de la sociedad de la información se ha generado una evolución sin precedente en el diseño e implementación de nuevas tecnologías informáticas y telemáticas, surgiendo a la par la necesidad de diseñar nuevas formas, como los lenguajes artificiales que permiten la comunicación hombre-máquina. Para Gelbukh [1], uno de los bienes más preciados de la humanidad es el conocimiento, los libros son un registro del mismo y se encuentran escritos en un sinfín de idiomas. Actualmente estos se almacenan en formato digital, así los computadores ayudan y optimizan el proceso de almacenamiento de conocimiento. Sin embargo, para un computador el conocimiento humano no pasa de ser un simple archivo o una dirección de memoria física, de esta manera lo que es conocimiento para los seres humanos, es una secuencia de señales digitales para las máquinas. El esfuerzo que la Ciencia invierte hoy en día para contrarrestar esta situación se denomina: procesamiento de lenguaje natural, procesamiento de texto, tecnologías de lenguaje o lingüística computacional. El lenguaje humano es complejo, multiforme y rico en expresiones, pero a la vez puede ser ambiguo y requerir interpretación de acuerdo al contexto, al momento y la intensión del hablante. El

[1] Ingeniero de sistemas, especialista en docencia universitaria, magister en educación y estudiante de doctorado en Ciencias de la educación. Docente de la UNAD y director del grupo de investigación GRIEE
[2] Ingeniera de Sistemas, magister en educación y docente de la Institución Universitaria CESMAG

lenguaje artificial, por el contrario, se diseña con una finalidad concreta, es restringido en su sintaxis y en su semántica, por ello es más preciso, con menos espacio para la libre interpretación y libre del contexto.

La investigación en el área del tratamiento computacional del lenguaje tiene como objetivo conseguir que el computador procese el texto por su sentido y no simplemente como un archivo binario. Actualmente, en el esquema general de la mayoría de los sistemas y métodos que involucran el procesamiento de lenguaje, el texto no se procesa directamente sino que se transforma en una representación formal que preserva sus características relevantes para procesamiento específico, como pueden ser: un conjunto de cadenas de letras, una tabla de base de datos o un conjunto de predicados lógicos. Posteriormente, el programa manipula esta representación y la transforma, buscando en ella las subestructuras necesarias; finalmente, los cambios hechos a la representación formal o la respuesta generada, se transforman nuevamente al lenguaje natural. [1]

En este orden de ideas, teniendo en cuenta las exigencias propias de la sociedad de la información, el auge de la computación, la informática, el potencial de las nuevas herramientas tecnológicas y los cambios en la cultura de las nuevas generaciones; es apremiante la investigación en el área de la inteligencia artificial, particularmente en cuanto al procesamiento automático del lenguaje natural. Temas complejos, pero a la vez de una gran expectativa social.

II. LENGUAJE NATURAL Y LENGUAJES ARTIFICIALES

La comunicación es el proceso por el cual dos o más individuos comparten un mensaje. En palabras de Fernández y Dahnke [2] es poner en común o compartir conocimiento, información, idea o actitud, haciendo uso de un medio y un código. La comunicación se lleva a cabo mediante señales, como: sonidos, gestos y símbolos.

Una señal se define como una realidad física que tiene un significado y a la asociación mental de la señal con su significado se conoce como signo. Por lo tanto, un signo tiene dos partes: la señal y el mensaje que la señal transmite. En lingüística, a la señal se la denomina significante y a aquello que ésta representa, significado [3]. Para que exista un signo es necesario que haya dos seres que previamente hayan acordado la asociación entre significante y significado, pues de nada serviría la luz roja del semáforo si los conductores no asociaran la luz con la prohibición de pasar.

Ahora bien, al conjunto de sonidos, señales y símbolos que los seres humanos utilizan para comunicarse, sea verbal o no, se le denomina lenguaje natural [4], porque para el ser humano, éste es una facultad natural. Cabe anotar que tomado en su conjunto, el lenguaje, es multiforme, heteróclito, físico, fisiológico, psíquico, y forma parte tanto del dominio de lo individual como de lo social.

De manera que se conoce como lenguaje natural al sistema que utiliza una comunidad lingüística con el fin primario de la comunicación, construido con reglas, convenciones lingüísticas y sociales durante el período de constitución histórica de esta sociedad. Según López [5], se denomina natural porque se adquiere de manera inconsciente, a diferencia de los lenguajes artificiales, que se aprenden por decisión o necesidad explícita, como es el caso de los lenguajes de programación.

Por otra parte, un lenguaje artificial es una definición de un vocabulario restringido proveniente del lenguaje humano, orientado por un conjunto de reglas que permiten construir expresiones aplicando: sintaxis, gramáticas y semántica determinadas. La definición de un lenguaje artificial obedece a una necesidad particular de comunicación, donde no es posible utilizar todo el potencial expresivo del lenguaje natural.

La definición de un lenguaje artificial tiene como finalidad evitar los inconvenientes de ambigüedad o vaguedad del lenguaje natural, por ello presenta un grado de artificialidad y convencionalidad mucho mayor por lo que se refiere a la construcción de símbolos y al significado que se les asigna. Ni los símbolos ni significados pertenecen a ninguna comunidad natural de hablantes, sino a grupos de hablantes relacionados por objetivos científicos o técnicos.

El caso más importante de lenguaje artificial es la definición de sistemas de codificación al mismo tiempo que de comunicación para la operación de computadores y comunicación entre expertos de esta área, entre ellos desempeñan un papel preponderante los lenguajes de programación del mismo modo que los protocolos de comunicaciones.

Los lenguajes de programación, al igual que el lenguaje natural, están diseñados para la comunicación de ideas entre personas, pero difieren en dos aspectos: el primero, tienen un dominio expresivo limitado ya que sólo permiten la comunicación de ideas algorítmicas o la definición de operaciones computacionales, cada lenguaje tiene su propia notación , además de que toda construcción debe regirse a ella; segundo, permiten la comunicación de ideas y significados entre personas, pero también entre personas y equipos de cómputo, pues están relacionados con las operaciones incorporadas en los microprocesadores [6].

Para que un computador realice alguna tarea es necesario proporcionarle una serie de instrucciones o una especificación, a la que comúnmente se conoce como programa. En los inicios de la computación, el único lenguaje para comunicarse con el computador era el lenguaje de máquina, cuyo vocabulario estaba conformado por códigos numéricos que representaban las operaciones

que la máquina podía realizar, con la enorme dificultad de que cada máquina tenía su propio sistema de códigos; es decir, era necesario aprender a comunicarse con cada máquina en particular [7].

Por ejemplo, el PC (*personal computer*) de IBM (*International Business Machines*) utilizaba un procesador Intel 8x86, para decirle a este equipo que colocara el número 2 en la posición de memoria 0000, había que escribir: *C7 06 0000 0002*.

El lenguaje de máquina fue reemplazado por el lenguaje ensamblador, en éste tanto las operaciones como las localidades de memoria se representan de forma simbólica. El equivalente para la instrucción anterior, en lenguaje ensamblador fue: *Mov x, 2*. Donde se asocia la letra *x* como identificador de la posición de memoria y la palabra *Mov* como la operación mover a.

El lenguaje ensamblador mejoró considerablemente la velocidad y exactitud con que pueden escribirse los programas y, aunque todavía se utiliza, se caracteriza por la dificultad para escribir, leer y comprender el código, así como por la dependencia de la máquina.

El siguiente paso en la evolución de los lenguajes de programación fue desarrollar un lenguaje donde las operaciones tuvieran una notación similar al lenguaje humano, ya sea al matemático o al natural, y que fuera independiente de la máquina, de manera que el mismo programa pudiera ser entendido por diferentes computadores, así aparecen los lenguajes de programación de alto nivel, donde la expresión que se ha utilizado como ejemplo puede escribirse: $x = 2$.

Los lenguajes modernos están especificados por un conjunto de reglas para construir instrucciones correctas y definidos semánticamente, es decir, describen de manera precisa lo que significa cada construcción en particular. De tal modo, las expresiones pueden ser organizadas en módulos y pasadas a un compilador el cual traduce el código en un lenguaje comprensible por una máquina en particular y finalmente contar con un programa.

III. PROCESAMIENTO COMPUTACIONAL DEL LENGUAJE NATURAL

El lenguaje natural se considera un instrumento sumamente adaptado a la comunicación de la vida ordinaria, pero ambiguo y vago desde el punto de vista informático. Mientras que los lenguajes formales, como los de programación, se caracterizan por ser concisos, no ambiguos, con una sintaxis estricta y una semántica concreta. Esto no significa que el lenguaje natural sea inferior a los lenguajes informáticos; por el contrario, su capacidad comunicativa es ilimitada, evolutiva, adaptativa y abierta a multitud de interpretaciones, características éstas a las que no aspiran los lenguajes artificiales.

El procesamiento del lenguaje natural se aborda desde la Lingüística Computacional, área de la lingüística aplicada a la Inteligencia Artificial cuyo objetivo principal es la realización de estudios informáticos que simulen la capacidad humana de hablar y entender [5].

En el lenguaje intervienen diversos factores cognitivos y psicológicos, sin tener que representar toda la estructura mental y cognitiva humana. Cada programa informático, según sea su función, se ocupará de unos u otros aspectos del lenguaje y sus estrechas relaciones con los demás componentes cognitivos, de manera que el trabajo se hace modular [8].

Los programas que escuchan un texto y lo escriben, como: traductores, software de reconocimiento de voz y síntesis de voz son las áreas más conocidas de la Lingüística computacional.

El Procesamiento del Lenguaje Natural (PLN) es una rama muy importante de la Inteligencia Artificial y una de las más antiguas, las primeras traducciones automáticas iniciaron en la década de los 40`s a la par que la II Guerra mundial, sin embargo a causa de la escasa potencia computacional los intentos fracasaron, pero a pesar de ello, a partir de la década del 60 el PLN resurgió nuevamente [9].

El punto de partida para el procesamiento del lenguaje natural es el análisis sintáctico. Éste es el encargado de realizar la verificación de las distintas reglas de formación de un lenguaje y de generar, como resultado de este proceso, representaciones gráficas en forma de estructura jerárquica o árbol sintáctico. Por medio de estos árboles se define si una expresión pertenece o no a un lenguaje [10]. Por su parte Zapata y Hernández [11] consideran que en la revisión de las reglas de producción, las cuales definen un lenguaje, y la verificación de la frase que se analiza para que cumpla con estas reglas de formación, realizada por un ser humano, tiene lugar la subjetividad en la interpretación y la posibilidad de generar problemas, de interpretación o de incompletitud en la información, aumentaría a medida que incrementa la masa de los elementos a analizar.

IV. APLICACIONES DEL PROCESAMIENTO DEL LENGUAJE NATURAL

Entre las múltiples aplicaciones que puede tener el procesamiento del lenguaje natural por medio de sistemas computacionales, se destacan:

Buscadores Automáticos. En Internet existe infinidad de documentos que contienen todo tipo de información, falsa y verdadera. La información no se encuentra bien distribuida en la red y encontrarla se hace difícil.

El algoritmo general de búsqueda es:

1. Búsqueda primaria de información

2. Lectura y comparación
3. Descartar documentos

Los buscadores tradicionales aún no entienden completamente qué necesita el usuario a pesar de haberse incorporado técnicas como el reconocimiento semántico. Mediante la implementación de las técnicas semánticas las búsquedas de información en Internet hoy en día son superiores a las de tres años atrás, de manera tal que el buscador parece comprender más de lo que el usuario "humano" necesita. Entre los proyectos más conocidos hasta la fecha se encuentran:

W3C *Semantic Web Activity*: Esta es una organización que mediante un esfuerzo colaborativo entre diferentes investigadores analizan y desarrollan nuevas técnicas de Web Semántica.

La traducción automática. Si en la Torre de Babel, los habitantes hubiesen tenido un traductor automático, habrían llegado al cielo. Según Gavaldà [12], la propuesta de la traducción automática es sencilla, consiste en transformar texto de una lengua a otra manteniendo el significado. El modelo más sencillo es el de la traducción léxica, que consiste en la sustitución de cada palabra por la correspondiente en la lengua a traducir. No obstante, esta estrategia no funciona satisfactoriamente por dos razones:

a. La correspondencia entre dos lenguas no es biyectiva.
b. La sintaxis o el orden de los constituyentes gramaticales es diferente en cada lengua.

Un nivel de complejidad mayor en cuanto a traducción se muestra, por ejemplo, con el traductor de Google, este realiza un reconocimiento léxico, comparación con reglas sintáctica según la probabilidad de aparición y finalmente hace una correspondencia entre parejas de palabras en los dos idiomas.

Un ejemplo conocido de la aplicación de esta tecnología fue la incorporación de software de traducción automática a teléfonos Smartphone por parte del ejército de Estados Unidos, durante la guerra en Afganistán a causa de la escasez de intérpretes y la falta de colaboración de los pocos conocidos.

La síntesis de la voz es la lectura automatizada en voz alta de un texto. Es decir, la transformación de una secuencia de palabras en formato digital a una señal acústica suficientemente comprensible como para que un hablante de una lengua sea capaz de recuperar el texto original [12].

La idea de dictar órdenes a un computador y que este las entienda parece de fantasía, sin embargo los proyectos de hoy en día diseñan tanto dispositivos como programas de reconocimiento de voz para que los computadores escuchen y ejecuten. Entonces, para copiar un texto ya no será necesario seleccionarlo, copiarlo y pegarlo, únicamente se debe dar una orden.

Uno de los proyectos más prometedores en ésta área es el Proyecto Debian, de software libre, que incorpora un software sintetizador de voz a partir de la entrada por micrófono.

Compresión del lenguaje natural: cuando la tarea es únicamente transcripción, como es el caso de dictar una carta, los procesos que el computador ejecuta son relativamente sencillos, pero si el nivel de complejidad incrementa requiriendo por ejemplo que las letras cambien de color o tipo de fuente implica que el computador entienda y ejecute; dando lugar al denominado procesamiento y compresión del lenguaje natural.

Desde el sentido absolutamente lingüístico es posible la creación de una máquina que hable y entienda. Esto significa que es posible generar infinitas frases coherentes y entendibles para los hablantes. Por ejemplo, se puede crear un programa que genere oraciones bien formadas sintácticamente sin tener en cuenta el significado de éstas, para lograrlo haría falta un lexicón y un conjunto de reglas combinatorias. Un lexicón se define como una lista de palabras que están almacenadas en el cerebro de cada hablante, las cuales se relacionan de manera compleja con sus respectivos significados, las reglas que permiten combinar el lexicón se denomina sintaxis, de esta manera una persona crea infinitud de oraciones [5].

El software se concentraría en las reglas sintácticas y el lexicón, puesto que relacionar la semántica es demasiado complejo, aunque existen ramas de la Inteligencia Artificial como: gramática independiente de contexto, gramática de cláusulas definidas, modelos probabilísticos y gramáticas probabilísticas, entre otros, que investigan y trabajan sobre soluciones computacionales para la comunicación hombre-máquina-hombre a través del lenguaje natural.

Por su parte, Gavaldà [12] afirma que el procesamiento del lenguaje en un computador se hace calculando la frecuencia de aparición de las letras mediante el cálculo matemático complejo, denominado: cálculo de probabilidades con n-gramas. Cada letra tiene una frecuencia de aparición diferente en cada idioma, en ruso por ejemplo, las r son más frecuentes; en español la letra ñ es característica y sobre la comprensión escribe:

Por compresión del lenguaje natural se entiende la transformación del texto en una representación semántica apta para el razonamiento y la ejecución de las órdenes. Esta representación se consigue a través del proceso de *parsing* o construcción de un árbol de análisis a partir de una gramática. Si la gramática es sintáctica, el árbol de análisis ofrece una información sobre las categorías gramaticales de las palabras y su función sintáctica, como la identificación del sujeto, verbo, predicado,

complementos, etc. Mientras que si la gramática es semántica, el árbol de análisis ya es bastante próximo a la representación lógica que permite el razonamiento y la ejecución.

El procesamiento del lenguaje natural es la aplicación tecnológica más deseada por sus aplicaciones prácticas, aparentemente es arte de ilusión que un computador transforme la señal acústica de la voz en una secuencia de dígitos binarios susceptibles de ser procesados computacionalmente.

El procesamiento digital del lenguaje natural en su forma verbal promete una revolución en la informática. Actualmente se cuenta con aplicaciones que procesan la señal de audio y la convierten en texto de manera que el computador puede escribir un dictado.

Se dispone también de sistemas que atienden llamadas telefónicas y mantienen un diálogo para proporcionar la información que el usuario requiere, se usa frecuentemente para conocer horarios de aviones o trenes, para encontrar un alojamiento en una ciudad e incluso para hacer las reservas. Esto implica una tecnología mucho más compleja que debe incluir el reconocimiento de la voz, la comprensión del lenguaje natural, el acceso a bases de datos, la síntesis de voz y un procesador de diálogo que sea el encargado de unir fluidamente todos los componentes [5].

Otras aplicaciones más avanzadas y que se espera estén disponibles en un futuro próximo son: las interfaces de audio, las cuales evitarían que el usuario tenga que usar el teclado y el ratón para interactuar con el computador, pues podría hacerlo de forma verbal, como se lo hace entre personas, esto no sólo agilizaría significativamente el trabajo, sino que también facilitaría el acceso a la tecnología a personas con limitaciones físicas; y los lenguajes de programación en notación verbal, esta tecnología sería consecuencia de la anterior y consistiría en construir los programas, ya no como secuencias de código escrito en lenguaje de programación, sino como especificaciones verbales haciendo uso de un conjunto de palabras previamente definidas en el lenguaje. El compilador comprenderá la semántica de las expresiones verbales y construirá el programa equivalente en lenguaje de máquina.

V. CONCLUSIONES

La comunicación es la facultad, que sin ser exclusiva de los seres humanos, es en estos donde se encuentra su mayor nivel de desarrollo, en la medida en que ésta ha evolucionado, la sociedad ha alcanzado mayores cuotas de progreso, correspondiéndole, en este orden de ideas, el crédito por ser la que ha posibilitado el compartir conocimientos, ideas, sentimientos y el desarrollo de proyectos mancomunados que han incrementado la ciencia y la tecnología.

El lenguaje humano es una rica colección de símbolos, sonidos y significados que hacen posible la comunicación. Éste se desarrolla progresivamente por los individuos de una sociedad en cuanto van apropiándose de las palabras, gestos y significados que utilizan en su vida diaria. Las palabras son memorizadas y luego se combinan produciendo frases nuevas con significados socialmente comprensibles, de ésta manera el lenguaje es versátil, heteróclito y adaptativo.

Las nociones del lenguaje natural se han llevado a las ciencias de la computación para la construcción de lenguajes y protocolos que permitan la comunicación entre el hombre y la máquina o entre máquinas. No obstante, es imposible crear sistemas capaces de procesar el amplio espectro de posibilidades que éste ofrece, por ello se han creado lenguajes artificiales de alto nivel, que son fragmentos selectivos del lenguaje natural, basados en reglas muy estrictas de sintaxis y con semánticas bien definidas que permiten la comunicación precisa y limitada al ámbito para el cual han sido diseñados tales lenguajes.

La inteligencia artificial está interesada en desarrollar aplicaciones capaces de procesar el lenguaje natural y en la actualidad se cuenta con avances importantes y proyectos ambiciosos, como los buscadores, traductores automáticos y lo sintetizadores de voz. Se está trabajando en aplicaciones semánticas para la Web y en interfaces para el lenguaje natural, siendo estas áreas, campos prometedores para la investigación y el desarrollo de nuevas tecnologías.

REFERENCIAS

[1] A. Gelbukh, *Tendencias recientes en el procesamiento de lenguaje natural*. Proc. SICOM-2002, Villahermosa, Tabasco, México. 2002.

[2] C. Fernández y G. Dahnke, *La Comunicación humana*. McGraw-Hill, México. 1994.

[3] M. Seco, *Gramática esencial del español*. Espasa Calpe, Madrid. 2004.

[4] Espasa Calpe. *Gran Enciclopedia Espasa*. Tomo 12. Bogotá. 2005.

[5] X. López, (2010) ¿Qué es la Lingüística Computacional o PLN?, 2004. website. [Online]. Available: http://www.aucel.com/pln/k-es.html.

[6] A. Tocker, y R. Noonan, *Lenguajes de programación: principios y paradigmas*. McGraw-Hill, Madrid, 2003.

[7] K. Louden, *Construcción de compiladores*. Thomson, México, 2004.

[8] Timarán et al. Un nuevo enfoque en la enseñanza de la programación, Universidad de Nariño, San Juan de Pasto, 2009.

[9] J. Carbonell, El procesamiento del lenguaje natural, tecnología en transición. Congreso de la Lengua Española, Sevilla, 1992.

[10] R. Mitkov, The Oxford Handbook of Computational Linguistics. Oxford University Press, New York, 2003.

[11] C. Zapata, y J. Hernández, "Analizador Sintáctico de Lenguaje Natural con Reglas Editables para la Generación de Primitivas UML". En: Revista Avances en Sistemas e Informática, vol.4 No. 1 Junio de 2007, Medellín, 2007. [Online]. Available: http://pisis.unalmed.edu.co/avances/archivos/ediciones/Edicion%20Avances%202007%201/1.pdf. Fecha de consulta: septiembre 10, 2010.

[12] Gavaldà, Marsal. Gavaldà, Marsal. (2010) La investigación en tecnologías de la lengua. [Online]. Available: http://www.prbb.org/quark/19/019021.htm, fecha de consulta marzo de 2010.