ARTÍCULO INVITADO

Procesamiento de Lenguaje Natural y sus Aplicaciones

por **Alexander Gelbukh**

Un cuento de una máquina parlante

En los cuentos para niños, los animales y las cosas inanimadas pero mágicas, se comportan como personas: inteligentemente. 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. El hombre siempre ha asociado la inteligencia con el habla.

En nuestros días la ciencia convierte cada vez más cuentos en una realidad. Ya no nos sorprende una alfombra voladora (aunque no parezca una alfombra) y ¿qué falta para que podamos conversar con Pinocho? En los números anteriores de **Komputer Sapiens** se ha hablado 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 ahora ha sido exclusivo de los humanos (y, claro, de las cosas mágicas).

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; así, una contestadora telefónica común, una impresora o un procesador de palabras como Microsoft Word tampoco son dispositivos o software de PLN, mientras que un traductor automático sin duda lo es.

Diferentes programas exhiben diferente grado del procesamiento inteligente del lenguaje. Por ejemplo, un buscador de documentos puede simplemente buscar los documentos que contienen la cadena de letras especificada por el usuario, sin importar que esta cadena tenga o no un significado en un lenguaje (como el español o el inglés). En este caso no sería una aplicación del PLN. Sin embargo, el mismo buscador podría buscar los documentos que comuniquen la idea especificada por el usuario, sin importar con qué letras la comunican, y 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.

La ciencia que estudia el PLN se llama lingüística computacional. El nombre fue inventado en los tiempos cuando eso era: lingüística para las computadoras. Los lingüistas, a través de la introspección e intuición, escribían las reglas y los diccionarios cada vez más exactos y detallados, acercándose al objetivo: dotar a la computadora con la capacidad de entender el lenguaje humano. Este camino era muy difícil y laborioso, y los avances, aunque impresionantes, eran lentos y esporádicos.

Todo eso cambió con la llegada de Internet. Los investigadores obtuvieron acceso a volúmenes gigantescos de textos, el objeto del estudio de nuestra ciencia, y esta última, en lugar de introspección e intuición, se convirtió en el estudio estadístico directo de los datos disponibles. La lingüística computacional, en su etapa actual de desarrollo, es principalmente una rama de las tecnologías de aprendizaje automático, una parte de la inteligencia artificial y la estadística.

El aprendizaje automático se dedica al descubrimiento totalmente automático de las regularidades y las relaciones en los datos. Usualmente se aplica a datos numéricos, pero la lingüística computacional puede ser considerada como el aprendizaje automático sobre un tipo de datos especial, los textos en un lenguaje humano. Es así como un niño aprende su lenguaje natal: nadie le enseña las reglas, las gramáticas y los diccionarios; en su lugar su cerebro analiza estadísticamente los sonidos del lenguaje y su relación con el medio ambiente, y aprende a reaccionar adecuadamente.

El PLN tiene un gran número de aplicaciones prácticas. Aunque el gran sueño de los investigadores es poder algún día conversar en viva voz con Pinocho (de lo cual quizá estamos menos lejos de lo que parece), avances incluso muy pequeños e insignificantes en comparación con este sueño llevan a grandes logros tecnológicos en las aplicaciones de las tecnologías del PLN.

Uso eficiente de nuestro tesoro: Búsqueda y presentación del texto

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. Cuando se trata de dinero, lo guardamos de tal manera para encontrarlo rápidamente cuando lo necesitamos y procuramos que no pierda valor con el tiempo. Pero cuando se trata de nuestro mayor tesoro, el conocimiento, lo manejamos de manera tan negligente como nunca hacemos con el dinero.

El conocimiento se almacena y se transmite en forma de lenguaje humano, los textos escritos, por ejemplo, en español o inglés. Sin embargo, en la actualidad usamos estos textos muy ineficientemente. Mencionaré cua-

 ISSN 2007-0691

tro componentes necesarios para su uso eficiente: la digitalización, la búsqueda, la presentación de la información y su uso directo por el software.

Primero, la digitalización de los documentos. Las bibliotecas tienen toneladas de libros en papel. Los archivos, tales como el Archivo General de la Nación, tienen 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.

Por digitalización aquí entiendo la obtención del texto como una secuencia de letras, no como una fotografía digital. Esto requiere de gran fuerza de PLN. 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.

Segundo, la búsqueda de la información relevante, llamada también recuperación de información. No sirve de nada el conocimiento escrito y guardado si no puede encontrarse cuando se necesita. El problema de la búsqueda es que la misma idea se puede expresar con muy diferentes palabras. Por ejemplo, el usuario expresa su interés con la frase "la derrota de Maximiliano" y el documento relevante para tal petición es "la victoria de Juárez". Los dos textos no tienen ninguna palabra en común, pero un humano, usando su experiencia lingüística (derrota—victoria) y su conocimiento del mundo (Maximiliano—Juárez) fácilmente detectaría la relevancia del documento para la petición.

Progresos muy significativos se han logrado para que los programas puedan utilizar este tipo de razonamiento para satisfacer de la mejor manera las necesidades de los usuarios.

Tercero, la presentación eficiente de la información contenida en los textos. El ejemplo más directo de esta tecnología es la construcción automática de resúmenes: dado un texto largo (o un millón de textos), un generador automático de resúmenes trata de detectar lo más importante que se comunica y presentarlo en un texto corto que se podrá leer en un tiempo razonable. A pesar del mucho esfuerzo que se ha dedicado a estas tecnologías, los resultados obtenidos hasta ahora son aún modestos, aunque cada vez mejores.

Otra manera de resumir la información contenida en muchos documentos y hacerlos más manejables es agruparlos y clasificarlos; en lugar de tener que leer millones de archivos, el usuario sólo necesitará considerar, digamos, cinco grupos cuyos documentos se parecen entre sí. O bien, diferentes personas considerarán cada grupo de documentos. Por ejemplo, en un gobierno, alcaldía o en una empresa grande, las quejas y peticiones de los ciudadanos o los clientes se dirigirán a las oficinas correspondientes.

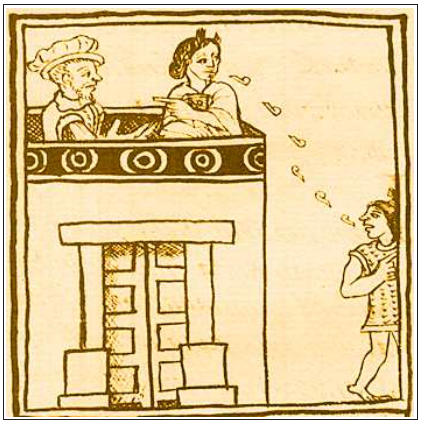
El resumen de la información relevante puede llegar a ser tan corto como una sola palabra. Es el caso de la respuesta automática a preguntas. ¿Para qué busca los documentos el usuario de un sistema de recuperación de información? Quizá no necesita los documentos sino tiene una duda e intenta aclararla leyéndolos. Las tecnologías de respuesta automática a preguntas lo hacen directamente: a la petición "¿Dónde nació Juárez?" la respuesta será "¡en Guelatao!" y no la biografía completa de Juárez. Tales sistemas se basan en un razonamiento complejo que a veces requiere de profunda comprensión del significado del texto: por ejemplo, pueden inferir la información requerida del texto "llegamos a Guelatao, el pueblo natal del Benemérito de Las Américas".

Otras maneras de resumir el contenido de muchos 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).

Cuarto, el uso de la información contenida en los textos por el mismo software para resolver tareas más complejas. La máquina puede encontrar el conocimiento necesario de los textos disponibles, tales como los artículos científicos o los libros de texto. Tales aplicaciones están actualmente en la fase experimental, aunque en el futuro se convertirán en la manera principal del manejo de conocimiento.

¿Cómo sabemos que un animal o una cosa son inteligentes? Porque son parlantes: hablan y entienden lo que les dicen. El hombre siempre ha asociado la inteligencia con el habla.

 ISSN 2007-0691



"Mujer azteca hablando" (superior) y "Malinche traduciendo" (inferior), detalles del *Código Florentino*, libro 12, capítulo 18 (1580).

Entender el lenguaje ajeno es la paz: Traducción automática

Parafraseando la célebre frase: 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, el PLN aporta las tecnologías de la traducción automática. Con esta tecnología el usuario puede leer en su propio lenguaje un texto escrito en otro lenguaje, puede escribir dirigiéndose a los lectores que hablan otros lenguajes o conversar (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. El traductor de Google, **www.google.com.mx/language_tools?hl=es**, nos permite sin ayuda externa leer las páginas de Internet en chino, árabe, ruso y muchas otras lenguas, sin mencionar el inglés. Sin embargo, mientras que el texto producido por tales traductores es muy útil y sirve de gran ayuda, es todavía muy mejorable. Estos sistemas son actualmente deficientes en dos aspectos principales.

Primero, la calidad del texto que producen. En muchas ocasiones parece haber sido escrito por un extranjero que no habla bien el español, y en otras de plano nos reprobarían en la primaria si escribiéramos así. El mejorar este aspecto requiere de mucho esfuerzo, pero es manejable y aunque a veces el texto se ve raro, no presenta tanta molestia en la práctica.

Segundo, y mucho más peligroso, 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 tener consecuencias graves por la generación de posibles malos entendidos e información falsa. Sin embargo, 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 requiere de toda la fuerza de la ciencia del PLN. En muchos casos es indispensable que el programa entienda el texto lo suficientemente bien para poder razonar sobre él. Con justa razón, la traducción automática desde el mismo comienzo del PLN fue su principal motivación, y fuente de inspiración y retos.

A pesar de dichas dificultades vale la pena seguir trabajando en esta tarea, pues una vez resueltos los problemas técnicos, viviremos en un mundo sin fronteras lingüísticas, sin limitaciones que se nos imponen por no hablar el inglés (o el chino, o el español) y sin tanta división cultural y social derivada de estas limitaciones.

Para hablar con un vecino del continente que no hable nuestro idioma, simplemente prenderemos el celular que se encargará de traducir lo que le estamos diciendo y de traducirnos también su respuesta.

 ISSN 2007-0691

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.

Quiero decir, vivo yo, mis colegas ingenieros, mis estudiantes y seguramente usted, querido lector. Pero ¡qué poquitos somos quienes vivimos en la era informática! Cuando hablo con algunos de mis conocidos médicos, abogados, músicos, historiadores, choferes, obreros, escucho "pues . . . la computadora . . . y estas cosas . . . ¡no soy bueno en esto!" 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 (como la necesidad de aprender informática). Deben ser nuestras ayudantes naturales, fáciles de usar. Deben aprender nuestro lenguaje y no obligarnos a aprender el suyo.

Los robots ya son físicamente capaces de ser nuestros sirvientes y ayudantes en tareas cotidianas. Según el gobierno de Corea del Sur, cada familia coreana en el año 2020 tendrá un robot ayudante en la casa [1], tal como en siglos pasados era común tener sirvientes. Bill Gates, el líder de Microsoft, dice también que habrá un robot en cada hogar [2].

Pero para que un robot se convierta en un verdadero ayudante de casa tiene que entender nuestro lenguaje. Esto significará la era informática para todos, no sólo para los ingenieros.

Entre muchos problemas técnicos en este camino mencionaré aquí cuatro. Ninguno de ellos es inherente a la tarea de las interfaces humano-computadora, pero son aquí más evidentes y los retos que presentan son más difíciles que en otras tareas.

El primero es el procesamiento de habla. Varias veces dije que los programas de PLN procesan, clasifican, analizan el texto. Pero no debe ser estrictamente así. No hablamos en texto, hablamos en voz. Para lograr una interfaz eficiente, las máquinas deben entender el lenguaje hablado (aunque internamente lo transformen a texto para analizarlo).

El segundo es la conducción del diálogo, el cual presenta retos distintos de los de un texto normal (monólogo). Por ejemplo, en el diálogo se usan mucho las oraciones incompletas o hasta recortadas a una sola palabra (como "ajá", "pues"). Además, hay ciertas reglas de conducta en cuanto al cambio de los 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 eso no se puede; hay que decidir cuál parte vamos a expresar en la primera oración y cuál en la segunda (y peor aún, dividir la idea grande en pedacitos de tamaño de oración), cuál palabra va primero y cuál luego; con qué palabra se expresa la misma idea en diferentes contextos. En español, por ejemplo, dar atención se dice "prestar", dar una clase se dice "impartir", dar una carta se dice "entregar", dar una enfermedad se dice "contagiar".

Finalmente, el cuarto problema es relacionar las palabras con las acciones, objetos y circunstancias en la conversación. Un robot ayudante debe poder reaccionar adecuadamente a frases como "ve allá y tráeme aquello", relacionando el objeto y la dirección con el movimiento del dedo del usuario.

Igual como en el caso de otras aplicaciones, mientras los investigadores nos están acercando a lo que hoy se ve como ciencia ficción, existen actualmente aplicaciones prácticas y factibles de esta tecnología. Una aplicación práctica de las interfaces humano-computadora son las interfaces con las bases de datos. Normalmente las preguntas aún bastante sencillas, como ¿qué porcentaje de los alumnos del tercer semestre reprobaron dos materias?, implican programación en un lenguaje especializado de consulta a bases de datos llamado SQL. Mucho esfuerzo se ha dedicado durante décadas a que las máquinas puedan directamente entender las preguntas en su forma natural, proporcionando así el acceso a la información a los usuarios comunes sin la necesidad de un programador intermediario.

Un ejemplo de la aplicación práctica del reconocimiento de habla son los sistemas de dictado, los cuales permiten que se dicten textos (como este artículo) con un micrófono en lugar de escribirlos con el teclado. La miniaturización de los sistemas electrónicos aumentará la importancia de la comunicación en voz: será la única (y muy natural) manera de interactuar con un reloj de pulsera inteligente.

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

Por PNL se entiende la habilidad de una máquina para procesar la información comunicada, no sólo las letras o los sonidos del lenguaje.

 ISSN 2007-0691

Y mucho, mucho más . . .

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

Los sistemas de soporte para la composición de textos proporcionan ayuda al usuario para escribir 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 (muy útil en los celulares); proporcionan traducciones, sinónimos y explicaciones de las palabras o sugieren palabras según su descripción [3]. Pueden variar en complejidad desde muy simples (tales como la división de las palabras con guiones) hasta muy complejos, por ejemplo, la verificación lógica y factual del texto (en la frase "al salir de Francia, Juan visitó su capital Londres" 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 al estilo, lenguaje o exactitud. En la educación asistida por computadora los métodos del PLN ayudan a componer los cursos y a proporcionar al estudiante la información requerida.

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 enorme cantidad de datos experimentales reportados, por ejemplo, en la investigación de la interacción de los genes y las proteínas, resulta necesario el procesamiento automático de tales publicaciones ya que una persona ya no puede leer ni siquiera las más relevantes para su trabajo.

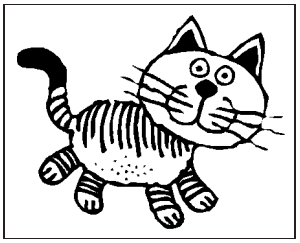
La lingüística forense aplica los métodos lingüísticos, y sobre todo computacionales, en las investigaciones criminalísticas y de peritaje. Estos métodos incluyen la identificación de la autoría de los textos o búsqueda de los fragmentos sospechosos en los mensajes o conversaciones grabadas. Dos áreas muy afines a la lingüística forense son la identificación de plagio (tanto en obras literarias o publicaciones científicas como en las composiciones de los estudiantes) y la esteganografía lingüística (los métodos para ocultar mensajes secretos en textos o habla y los métodos para detectar tales mensajes ocultos).

Las ideas y técnicas desarrolladas originalmente para el análisis del lenguaje resultan aplicables en áreas muy lejanas del lenguaje humano. Un ejemplo obvio es la teoría de compiladores y los lenguajes de programación, cuya creciente complejidad los aproxima cada vez más 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, dado que su autor es lingüista.

La genómica y la biología molecular comparten muchas ideas y métodos con el PLN, 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. Por razones similares, los métodos de PLN se emplean en el análisis y la generación automática de música: las estructuras repetitivas musicales se describen bien con las así llamadas gramáticas formales desarrolladas originalmente para la descripción de los fenómenos lingüísticos.

¿Qué gato tiene Juan?

Juan usa un gato para reparar su coche. ¿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 se 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.**

 ISSN 2007-0691

Por dónde continuar . . .

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 a los expertos del área. Si es usted un directivo o un empresario y encontró en este artículo algo que le puede servir, quizás se pregunte quién le puede dar el servicio, y si es usted estudiante (tal vez potencial, pues nunca es tarde para estudiar), quizás se pregunte en dónde puede obtener más información.

Al final de esta contribución se proporcionan enlaces a portales de asociaciones profesionales y a material disponible en línea. ¡Espero que les sean de utilidad!

INFORMACIÓN ADICIONAL

- El lector interesado puede encontrar más información y vínculos a las fuentes y los eventos relevantes en la página de la AMPLN, la Asociación Mexicana para el Procesamiento de Lenguaje Natural: **www.AMPLN.org**. La comunidad nacional del PLN organiza anualmente dos congresos con ponencias 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: **www.CICLing.org**, aunque no siempre en México.
- Para la lectura inicial se recomiendan los libros [4-6] disponibles desde la página **www.Gelbukh.com**, donde se puede también encontrar muchos artículos científicos sobre el tema y otros materiales relevantes.

REFERENCIAS

1. "A Robot in Every Home by 2020, South Korea Says", *National Geographic*, news.nationalgeographic.com/news/2006/09/060906-robots.html, visitado el 11 de febrero de 2010.
2. Gates B. (2007) "A Robot in Every Home", *Scientific American*, www.scientificamerican.com/article.cfm?id=a-robot-in-every-home, visitado el 11 de febrero de 2010.
3. Sierra G. (2001) "Búsqueda de palabras a partir de las definiciones en los diccionarios de lengua automatizados", *Actas de* 7^0 *Simposio Internacional de Comunicación Social*, 2, Santiago de Cuba.
4. Bolshakov I.A., Gelbukh A. (2004) *Computational linguistics: models, resources, applications*, IPN–UNAM-Fondo de Cultura Económica.
5. Gelbukh A., Sidorov G. (2010) *Procesamiento automático del español con enfoque en recursos léxicos grandes*, Segunda edición, ampliada y revisada. IPN.
6. Galicia Haro S.N., Gelbukh A. (2007) *Investigaciones en análisis sintáctico para el español*, IPN.

SOBRE EL AUTOR

Alexander Gelbukh es maestro en ciencias con especialidad en matemáticas y doctor en ciencias de la computación. Desde 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 II, 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.

 ISSN 2007-0691