La Biblioteca del ISCIII participa en un proyecto de inteligencia artificial para mejorar la búsqueda y manejo de información científica

| 10/05/2021 |

Elena Primo (directora), María Teresa Romera y Cristina Bojo, de la Biblioteca Nacional de Ciencias de la Salud del ISCIII.

El ISCIII, a través de su **Biblioteca Nacional de Ciencias de la Salud (BNCS)**, está participando en el proyecto europeo **MESINESP2** (Medical Semantic Indexing in Spanish Shared Task), que investiga el uso de inteligencia artificial aplicada a la minería de textos y semántica del lenguaje para facilitar y acelerar la búsqueda y manejo de información científica.

La rápida acumulación de publicaciones científicas relacionadas con biomedicina y ciencias de la salud hace cada día más necesario acceder de forma más rápida y efectiva a información biomédica y sanitaria. La búsqueda, manejo e interpretación de publicaciones médicas escritas en diversos idiomas se ha hecho más patente que nunca durante la pandemia de COVID-19, que exige un acceso global a fuentes fiables de información científica. La comunidad científica y sanitaria necesita a diario recuperar artículos biomédicos y sanitarios claves para desarrollar la medicina basada en evidencias, elaborar guías clínicas, llevar a cabo revisiones sistemáticas y publicar nuevos estudios científicos. Para hacerlo, los profesionales recurren a fuentes de información que, para ser útiles y efectivas, deben contar con un trabajo previo de Indización que mejore la recuperación de la información de interés que se ajuste al tema de búsqueda.

En este contexto, actualmente se está celebrando la iniciativa internacional MESINESP2 (Medical Semantic Indexing in Spanish Shared Task), en la que colabora el ISCIII. MESINESP2 es una competición internacional de indexación semántica de literatura científica, ensayos clínicos y patentes, en idioma

castellano, y se trata de un proyecto incluido en el **Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)**.

La unidad de Text Mining del **Barcelona Supercomputing Center (BSC-CNS)** impulsa, organiza y coordina en España esta iniciativa, en colaboración con la Biblioteca Nacional de Ciencias de la Salud (BNCS-ISCIII) y el Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud (**BIREME**), que trata de fomentar el desarrollo de sistemas de indexación semántica basada en los últimos avances de inteligencia artificial y procesamiento del lenguaje natural. En esta edición de MESINESP2 participan 35 grupos de diversos países (España es el que más aporta, con 9); tras cerrarse el plazo de envío de propuestas, el plazo de entrega de resultados comenzó el pasdo viernes 7 y está abierto hasta el próximo día 17 de mayo.

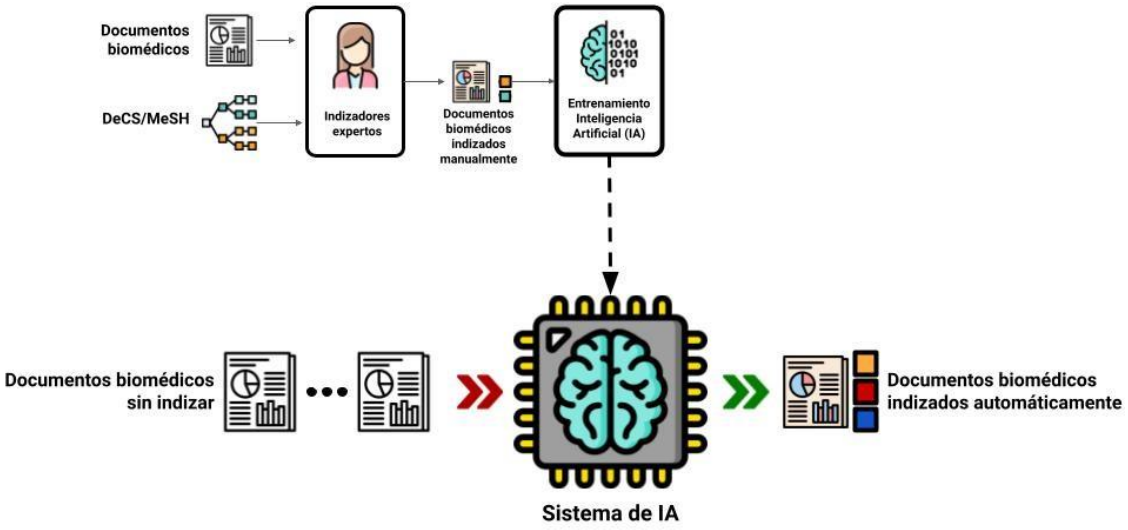
Definición de términos y representación de conceptos

Realizar búsquedas eficientes en estas fuentes de información requiere a menudo de consultas complejas, cuyo éxito depende en parte de la asignación previa, por parte de expertos, de términos específicos para describir su contenido. Este proceso, conocido como indización, sigue siendo en la actualidad una tarea manual, que debido al número de publicaciones en continuo crecimiento, es cada vez más inabarcable. Los términos que se utilizan en esta indización son términos controlados pertenecientes a lo que se denomina 'tesauros', listas de palabras o términos controlados de manera jerárquica que se emplean para representar los conceptos incluidos en un texto.

En el área de las ciencias de la salud, el tesauro más utilizado en inglés es el MeSH (Medical Subjects Headings) elaborado por la NLM (National Libtrary of medicine) de EE. UU. En español se utiliza el DeCS (Descriptores en Ciencias de la Salud), que es una traducción del MeSH impulsada por la OMS, que incluye algunas categorías suplementarias como Salud Publica,…y que se publica en versión trilingüe (Inglés, español y portugués). La **Biblioteca Virtual de Salud (BVS)**, una red internacional para gestión de la información y conocimiento en salud que en España coordina la BNCS-ISCIII, ofrece junto con bases de datos bibliográficas como **IBECS**, **LILACS** y **SciELO** acceso a una gran variedad de recursos de información científica en salud, y utiliza como herramienta para mejorar la recuperación los mencionados tesauros MeSH y DeCS.

Según explican **Elena Primo** y **Cristina Bojo**, de la Biblioteca Nacional de Ciencias de la Salud, los participantes de MESINESP2 "catalizarán la búsqueda de información biomédica a través de sistemas de indexación semántica basados en rigor científico y en las tecnologías más avanzadas de inteligencia artificial aplicadas a textos en español". Los sistemas que participan en esta iniciativa acelerarán la recuperación de información biomédica y sanitaria, facilitando la localización de textos relevantes en la literatura médica, así como sobre patentes y ensayos clínicos. Los resultados de MESINESP2 también generarán "sistemas de indexación semántica que serán potencialmente útiles para procesar otro tipo de contenido, tales como historia clínica electrónica, guías de práctica clínica o patentes", señalan.

Minería de datos e indización de textos



Infografía que representa el proceso de indización de textos científicos con ayuda de una herramienta de inteligencia artificial.

MESINESP2 forma parte de una línea de proyectos del Plan TL basados en el desarrollo de tareas colaborativas y competitivas ('shared tasks' o 'challenge tasks/tracks', en inglés). Cuando estas tareas están orientadas a participantes académicos, grupos de investigación y entorno empresarial, se articulan a través de campañas de evaluación de sistemas de Procesamiento de Lenguaje Natural (PLN) y minería de textos (proceso que permite buscar, extraer, analizar y derivar nueva información a partir de diversos textos o documentos).

Estas tareas y campañas de evaluación permiten evaluar de forma independiente, con métodos científicos y usando conjuntos de datos bien definidos, la calidad de los resultados obtenidos por los sistemas y algoritmos predictivos que participan en estos análisis. En concreto, MESINESP2 forma parte del **Proyecto BioASQ** de indexación de literatura biomédica.

La indexación (o indización) de literatura científica es una tarea documental que consiste en asignar a un documento los términos -procedentes de los tesauros- que describen, de forma unívoca, el contenido de un documento. El que una base de datos tenga sus registros indizados es un plus de calidad añadido, ya que permite al usuario realizar búsquedas a través de esos términos que, al ser controlados, evitan los accidentes propios del lenguaje natural, como las sinonimias y las polisemias.

Por ejemplo, en una base de datos con tesauro (como **Pubmed** o IBECS, desarrollada en el ISCIII) si uno quiere buscar trabajos sobre VIH, bastara con usar el descriptor aceptado para ello para que el sistema devuelva toda la información existente; si la base de datos no cuenta con ese tesauro, el usuario debería buscar por todos los posibles sinónimos: sida, HIV, VIH+, síndrome de

inmunodeficiencia adquirida, etc. y, aun así, no existiría la seguridad de haber localizado todos los datos.

Entrenamiento y aprendizaje tecnológico

La tarea de indización es, por su naturaleza, altamente especializada y compleja, requiere la lectura del documento y seleccionar los términos que describen ese contenido, por lo que también es costosa en tiempo. Por ello, el principal objetivo de MESINESP2 es construir una herramienta, basada en sistemas de **Procesamiento de Lenguaje Natural (PLN)**, que ayuden y mejoren la eficiencia de la indización manual, automatizando parte del proceso esta indización de forma automática.

Estos sistemas de PLN, basados en inteligencia artificial, requieren para su desarrollo de un 'entrenamiento' de la máquina, que debe 'aprender' a leer los textos con ayuda de un conjunto de fuentes, documentos y referencias. MESINESP2 obtiene toda esta información necesaria de la base de datos IBECS, mantenida por la BNCS-ISCIII, y de la base de datos LILACS, mantenida por BIREME.

Además de proporcionarle al sistema los textos, la máquina los necesita ya indizados para poder 'aprender' a ejecutar la tarea; es decir, necesita una guía precisa, un 'gold estándar' que, en este caso, también procede de IBECS y LILACS, las únicas bases de datos con literatura científica en español indizada con el tesauro DeCS. Por ello, la aportación de la BNCS-ISCIII a MESINESP2 es clave, ya que ha colaborado en la localización de corpus lingüísticos y científicos, ha proporcionado sus fuentes y bases de datos, ha asesorado en el proceso de indización y localizado casos de uso exitosos, entre otras cuestiones.