Sistemas de Acceso Inteligente a la Información Biomédica: una revisión

Biomedical Information Intelligent Access Systems: a review

Laura Plaza Morales (1), Jorge Carrillo de Albornoz Cuadrado (2), Juan Carlos Prados Frutos (3)

1. Becaria programa FPU. Departamento de Ingeniería del Software de la UCM
2. Investigador contratado programa IV PRICIT. Departamento de Ingeniería del Software de la UCM
3. Profesor Titular. Departamento de Anatomía y Embriología Humana I. Facultad de Medicina. Universidad Complutense de Madrid.

Correspondencia:
Laura Plaza Morales
Facultad de Informática. Universidad Complutense de Madrid
C\ Profesor José García Santesmases, s/n
28040, MADRID

Fecha de recepción: 22 junio 2009
Fecha de aceptación: 6 octubre 2009

Los autores declaran no tener ningún tipo de interés económico o comercial

Resumen

En un entorno como el de la medicina, caracterizado por la sobrecarga de trabajo y la escasez de tiempo, los sistemas inteligentes de acceso a la información pueden y deben utilizarse para facilitar la labor de investigadores y profesionales. Sin embargo, sorprende comprobar la escasa implantación de estos sistemas. Las razones son varias. En primer lugar, el potencial completo de estas estrategias sólo se alcanzará cuando la informática esté completamente integrada en la práctica médica. En segundo lugar, todavía es necesario avanzar en la estandarización de la estructura y el contenido de la información de los pacientes, así como en el uso de una terminología unificada y controlada. Aunque, especialmente durante la última década, los avances en ambos sentidos han sido considerables, lo cierto es que todavía queda mucho camino por recorrer.

Palabras clave: Acceso inteligente a la información, Procesamiento de Lenguaje Natural, Unified Medical Language System (UMLS), SNOMED, Medical Subject Headings (MeSH).

Abstract

Modern medical environment is characterized by the work overload and the lack of time. In such an environment, intelligent information access systems can undoubtedly assist the work of physicians and researchers. Nonetheless, despite their benefits, clinical information systems barely include these technologies. Several are the reasons. First, the potential of these strategies will only be achieved once health care professionals become familiar with them. Second, it is necessary to progress in the development of patient information standards, as well as in the use of an unified and controlled terminology. Even if important advances have been reached during the last decade, there is still much work to do.

Keywords: Intelligent information access, Natural Language Processing, Unified Medical Language System (UMLS), SNOMED, Medical Subject Headings (MeSH).

1. Introducción

Durante las últimas décadas, los avances en las Tecnologías de la Información y la Comunicación han alterado sustancialmente la naturaleza de la educación, la investigación y la práctica médica. La sobrecarga de información caracteriza a la medicina moderna, a la vez que la documentación disponible en formato electrónico crece y se vuelve más imprescindible tanto en la formación de los futuros médicos como en el ejercicio de la profesión. Esta información, en su mayor parte desestructurada y en lenguaje natural, puede provenir de muy diversas fuentes y presentar formatos y estilos de redacción muy dispares[1].

Según algunos estudios estadísticos[2], los médicos están suscritos a un promedio de 7 revistas, lo que significa que reciben alrededor de 2000 artículos nuevos cada año. Es obvio que tal cantidad de información hace imposible una revisión exhaustiva de la misma. Precisamente con el objetivo de facilitar el acceso eficiente a la información, durante los últimos 35 años investigadores en medicina e informática han aunado sus esfuerzos en aras de la definición y el desarrollo de sistemas de recuperación y tratamiento de información, y de apoyo a la toma de decisiones clínicas. Dejando a un lado los complejos sistemas de predicción y de análisis de decisiones, aún en un estado muy preliminar y con un elevado grado de incertidumbre, existen otros tipos de sistemas que, si bien no sustituyen ni aconsejan al médico en la toma de decisiones, facilitan en gran medida la labor de abstraer y recuperar en cada momento el conocimiento relevante dentro del océano de información disponible, e incluso dotar a la misma de cierta estructura que simplifique su análisis posterior[3,1].

Los primeros sistemas de este tipo, y quizás por ello los más populares, son los Sistemas de Recuperación de Información (*Information Retrieval*), cuyo propósito es encontrar documentos que respondan a las necesidades de información planteadas por el usuario. A partir de estos, surgen sistemas más sofisticados, como los de categorización (*Text Classification*), capaces de clasificar documentos (bien artículos científicos, bien historiales clínicos) en una serie de categorías predefinidas; o los de generación automática de resúmenes (*Automatic Summarization*) que, como su nombre indica, persiguen generar resúmenes que condensen la información importante de los documentos originales. Finalmente, en los últimos años se han presentado algunos trabajos que estudian la aplicación de métodos automáticos para la recuperación de historias clínicas similares a una dada.

En cualquier caso, todos estos sistemas presentan dos aspectos en común: por un lado, el uso de técnicas de Inteligencia Artificial (IA) y de Procesamiento de Lenguaje Natural (PLN); y por otro lado, la dificultad de trabajar con texto en un lenguaje específico, el lenguaje biomédico, cuyas peculiaridades condicionan el éxito de los sistemas de PLN tradicionales. Como se estudiará más adelante, características como la frecuente presencia de sinónimos y homónimos, o el uso de abreviaciones y neologismos, imponen la modificación de los algoritmos y métodos clásicos para tratar convenientemente estos fenómenos lingüísticos.

A pesar de las ventajas que presentan este tipo de sistemas, lo cierto es que hasta la fecha sólo han conseguido un éxito y un nivel de implantación limitado. Las razones son varias. En primer lugar, es obvio que el potencial completo de estas estrategias sólo se alcanzará cuando la informática esté completamente integrada en la práctica médica. En segundo lugar, todavía es necesario avanzar en la estandarización de la estructura y el contenido de la información de los pacientes, así como en el uso de una terminología unificada. De hecho, y como recogen la mayoría de las investigaciones más recientes, el punto de partida de estas tecnologías debe ser la detección de términos biomédicos en el texto y su traducción a conceptos de tesauros y ontologías del dominio, lo cual sólo será posible una vez se adopte un vocabulario común y controlado.

En relación a la información de los pacientes, la progresiva implantación de la Historia Clínica Electrónica (HCE) en nuestro país ha abierto el debate sobre la necesidad de abordar la adopción de estándares que definan, entre otros, los contenidos y la estructura del historial, la codificación de los datos clínicos y los mecanismos de confidencialidad y autentificación. En lo que respecta a las terminologías clínicas, merece la pena insistir en el papel fundamental que, durante los últimos años, han pasado a jugar en el desarrollo de sistemas de

información. En concreto, tres son las nomenclaturas más utilizadas: UMLS, MeSH y SNOMED-CT. Aunque se estudiarán en detalle a lo largo de este trabajo, cabe mencionar que la elección entre una u otras dependerá en gran medida de las necesidades concretas del sistema que se desea construir.

A lo largo de este trabajo se analizarán distintos tipos de sistemas de acceso a la información biomédica, incidiendo en su utilidad práctica y presentando las técnicas y métodos en los que se apoyan. Así mismo, se identificarán los principales problemas a los que se enfrentan y el importante papel que los estándares y las terminologías clínicas habrán de jugar en el éxito de estos sistemas.

2. Terminologías y Ontologías Biomédicas

Antes de avanzar en el estudio de los sistemas, resulta conveniente conocer qué se entiende por terminología y ontología en el dominio que nos ocupa, así como algunas de sus implementaciones más populares. Citando a Bodenreider [4], las ontologías biomédicas proveen un marco organizacional de los conceptos involucrados en entidades y procesos biológicos, en un sistema de relaciones jerárquicas y asociativas que permite razonar sobre el conocimiento del dominio. Las terminologías biomédicas, por su parte, promueven una manera estándar de nombrar los conceptos del dominio. Sin lugar a dudas, los recursos más utilizados en la recuperación de información médica son SNOMED-CT[5], UMLS[6] y MeSH[7], y es por ello que se ha considerado apropiado incluir una breve descripción de cada uno de ellos.

SNOMED-CT son las siglas de *Systematized Nomenclature of Medicine Clinical Terms*, una extensa terminología médica desarrollada por el *College of American Pathologists (CAP),* y mantenida por *The International Health Terminology Standards Development Organisation* (*IHTSDO*). Disponible en inglés, español y alemán, proporciona un lenguaje común que facilita la indexación, el almacenamiento, la recuperación y la agregación de datos médicos. La versión actual en español recoge más de 315.000 conceptos, 800.000 descripciones y 945.000 relaciones semánticas. Los componentes básicos de SNOMED son:

- **Conceptos**: representan una unidad mínima de significado.
- **Jerarquías**: compuestas por categorías de primer nivel, que a su vez se descomponen en sub categorías.
- **Relaciones**: que enlazan conceptos entre sí. Existen dos clases de relaciones: de tipo "es un", que conectan conceptos en una jerarquía; y las relaciones de atributos, que enlazan conceptos entre jerarquías.
- **Descripciones**: términos o nombres asociados a un concepto, que aportan una mayor flexibilidad en la expresión de los conceptos médicos.

UMLS (*Unified Medical Language System*), desarrollado por la *National Library of Medicine (NLM)* de los Estados Unidos, es un sistema que garantiza referencias cruzadas entre más de treinta vocabularios y clasificaciones, incluyendo la Clasificación Internacional de Enfermedades*, MESH* y *SNOMED.* UMLS presenta tres fuentes de conocimiento:

- El **Meta Tesauro** es una base de datos multilingüe que contiene información sobre conceptos biomédicos, incluyendo sus diferentes nombres y sus relaciones. Está construido a partir de las versiones electrónicas de diferentes tesauros, clasificaciones y listas de términos controlados utilizados en el cuidado de pacientes, en la elaboración de estadísticas sobre salud, en el indexado y la catalogación de literatura biomédica y en la investigación clínica. Está organizado por conceptos o significados. Su propósito es enlazar nombres alternativos y vistas de un mismo concepto, así como identificar relaciones útiles entre diferentes conceptos. Todos ellos están asignados al menos a un tipo de la *red semántica*. Muchas de las palabras y términos que aparecen en el meta tesauro también aparecen en el *léxico especializado*.
- El **Léxico Especializado**, en lengua inglesa, contiene en su versión actual unos 108.000 informes léxicos y más de 186.000 cadenas de términos. Cada entrada presenta información sintáctica, morfológica y ortográfica, incluyendo la categoría sintáctica (verbo, sustantivo, adjetivo, adverbio, pronombre…), las inflexiones de género y número, las conjugaciones de los verbos, los comparati-

vos y superlativos de los adjetivos y adverbios, e incluso posibles patrones de complementación (objetos y otros argumentos que pueden acompañar a los verbos, nombres y adjetivos).

- La **Red Semántica** presenta 132 tipos semánticos, y garantiza una categorización consistente de todos los conceptos representados en el meta tesauro. Los 53 enlaces entre tipos semánticos establecen la estructura de la red y representan las relaciones más importantes en el dominio biomédico. El enlace principal es el "es un", que establece una jerarquía entre los tipos semánticos, que se clasifican a su vez en seis agrupaciones básicas: *organismos, estructuras anatómicas, funciones biológicas, productos químicos, eventos, objetos físicos* y *conceptos o ideas*.

MeSH son las siglas de *Medical Subject Headings*, un tesauro desarrollado por la *National Library of Medicine* de los Estados Unidos consistente en un conjunto de términos, denominados descriptores (*descriptors)*, dispuestos en una estructura jerárquica que permite la búsqueda a varios niveles de especificidad. Los descriptores se organizan de dos modos distintos: alfabéticamente y mediante una estructura jerárquica de once niveles. En el primer nivel de la jerarquía se encuentran descriptores muy amplios, como *anatomía* o *desórdenes mentales,* mientras que conforme se desciende en la jerarquía, los descriptores se concretan, de manera que en el último nivel se encuentran conceptos como *tobillo.* En la versión 2008 de MeSH, se cuentan 24.767 descriptores, además de 172.000 conceptos suplementarios (*Supplementary Concept Records*) recogidos en un tesauro separado. Hay también más de 97.000 términos de ayuda para localizar el descriptor más apropiado. Los árboles de descriptores no constituyen una clasificación exhaustiva de las materias, sino que están diseñados para la búsqueda en la base de datos MEDLINE; además de como guía para las personas encargadas de asignar categorías a documentos.

3. Sistemas de Acceso a la Información Biomédica

El propósito de esta sección no es elaborar un catálogo de las tecnologías existentes para el acceso a la información médica, sino estudiar algunas de sus aplicaciones más destacadas con el objetivo de analizar las posibilidades que ofrecen las técnicas de Procesamiento de Lenguaje Natural en la construcción de este tipo de sistemas, así como los problemas a los que se enfrentan.

- Recuperación de información:

El término Recuperación de Información (RI) se utiliza para designar a un amplio abanico de técnicas cuyo objetivo es la búsqueda de documentos, principalmente en bases de datos o en la World Wide Web, y de la que los motores de búsqueda son el ejemplo más paradigmático. El proceso de búsqueda comienza cuando un usuario introduce una consulta en el sistema, expresión de sus necesidades de información, y concluye cuando éste devuelve una lista de los documentos que se ajustan a la consulta, ordenados de mayor a menor relevancia según el resultado proporcionado por una función de ranking.

Los investigadores y profesionales en medicina utilizan MEDLINE a diario para localizar artículos que satisfagan sus necesidades de información. Es por ello que la mayoría de las iniciativas en RI biomédica trabajan con esta base de datos.

La búsqueda booleana es el método más simple y extendido a día de hoy. Este tipo de búsqueda se basa en la ocurrencia o no, en los documentos de la base de datos, de las palabras "clave" introducidas por el usuario en la consulta, permitiendo además el uso de operadores lógicos como la conjunción, la disyunción y la negación[8]. Un ejemplo de sistema booleano es el utilizado en el motor de búsqueda de MEDLINE, que combina descriptores MeSH con palabras presentes en los *abstracts* y títulos de los artículos.

Por otra parte, un buen número de los trabajos tradicionales de RI utilizan un modelo del espacio vectorial[9]. En este modelo, la idea básica reside en la construcción de una matriz donde las filas (vectores) representan a los documentos de la base de datos, y las columnas corresponden a los términos incluidos en ellos.

De esta manera, cada documento se expresa como un vector *d=(1,0 ,0 ,2 ,3 , ..., 2)*, siendo cada uno de estos valores el número de veces que cada término aparece en el documento. La propia consulta se representa también como un vector de términos, y el cálculo de la similitud entre ésta y cada uno de los documentos de la base de datos se realiza, generalmente, utilizando la *función del coseno*, que equivale a calcular el producto escalar de los dos vectores. El resultado obtenido es un valor entre 0 (mínima similitud) y 1 (máxima similitud).

Finalmente, las investigaciones más recientes incorporan a los mecanismos tradicionales el uso de ontologías y otros recursos lingüísticos, con el objetivo de realizar búsquedas orientadas a conceptos (vs. términos). De este modo, tanto la consulta como los documentos de la base de datos se traducen a conceptos de una ontología, para posteriormente aplicar un modelo de espacio vectorial[10] u otros más complejos que utilizan grafos semánticos para capturar las relaciones entre los conceptos[11].

- Clasificación automática

La clasificación automática de documentos médicos (artículos, historiales clínicos, etc.) es otra de las áreas que está adquiriendo un gran auge en la actualidad. Estos sistemas centran su objetivo en la asignación automática de categorías de interés, definidas previamente por expertos, a un conjunto de documentos[12]. Esta operación se puede realizar bien sobre documentos enteros, bien sobre fragmentos de ellos, e incluso sobre palabras. Otro aspecto diferenciador es el grado de supervisión humana necesaria en el proceso de clasificación. Normalmente se utilizan conjuntos de entrenamiento, que no son más que colecciones de documentos previamente clasificados por expertos, para que los sistemas puedan aprender en base a ellos. Más concretamente, en los sistemas de clasificación se suelen utilizar dos conjuntos de entrenamiento: uno positivo, que contiene aquellos documentos que corresponden a cada categoría; y uno negativo, que agrupa a los documentos que en ningún caso pertenecen a la misma.

Cada año se generan en el mundo alrededor de dos millones de artículos nuevos en medicina[13], por lo que parece razonable la necesidad de disponer de sistemas capaces de clasificar automáticamente, en ciertas líneas de interés, dichos textos. Con ellos se consigue un mayor y mejor acceso del material generado, así como la posibilidad de crear y mejorar los sistemas de recuperación y búsqueda que hacen uso de las bases de datos. De nuevo, sirva de ejemplo la base bibliográfica MEDLINE, cuyo tamaño se incrementa exponencialmente año a año, y que actualmente cuenta con once millones de entradas. MEDLINE utiliza el tesauro MeSH para etiquetar los documentos; y en base a este etiquetado, se han desarrollado numerosos sistemas de clasificación automática, que actúan tanto sobre los documentos de MEDLINE como sobre los resultados de las búsquedas realizadas en la base de datos[14,15]. Actualmente, estos sistemas logran un porcentaje de acierto que varía entre el 60 y el 80 por ciento.

Como ya se ha comentado, los orígenes de esta disciplina se encuentran en la necesidad y el deseo de la correcta estructuración y diseño de las grandes bases de datos de información médica[8]. Las primeras aproximaciones se basaban en la ocurrencia o no de ciertas palabras claves asociadas a las categorías, y sobre las que en los últimos años se han aplicado diferentes métodos matemáticos[16] y de inteligencia artificial[17] para mejorar su efectividad. Otros enfoques resuelven el problema utilizando sistemas estadísticos y de probabilidad para determinar la categoría a asignar[18]. Una conclusión común a la que han llegado estos sistemas es, por ejemplo, que la información relevante para la categorización suele aparecer en el inicio del documento. Actualmente, la mayoría de sistemas hace uso de una combinación de ambas técnicas, junto con la utilización de ontologías y bases de conocimiento que permitan trabajar con los conceptos asociados a las palabras en lugar de con los términos en sí mismos.

- Codificación automática:

Aunque los sistemas de codificación de historiales clínicos pueden considerarse un caso particular de los sistemas de clasificación, su creciente demanda justifica el dedicarle un apartado propio.

Los sistemas de codificación, a diferencia de los de clasificación, persiguen asignar códigos a documentos, principalmente a las historias clínicas de los pacientes. A medida que la codificación de los HCE según la Clasificación Internacional de Enfermedades (ICD, Interna-

tional Classification of Diseases) se convierte en una práctica habitual en hospitales y centros sanitarios, y dada la dificultad que conlleva la asignación de estos códigos, el interés por desarrollar sistemas que realicen automáticamente esta tarea ha ido en aumento.

Precisamente con el objetivo de potenciar la investigación en este campo, durante los últimos años se han organizado distintas tareas competitivas, como el *Medical Natural Language Processing Challenge*[19], que propone la clasificación automática de historiales radiológicos en base a la codificación ICD-9-CM (*International Classification of Diseases, Ninth Revision, Clinical Modification*). Por regla general, estos sistemas utilizan métodos similares a los utilizados en la clasificación con categorías. Sin embargo, dado que estos códigos son mucho más específicos que las categorías utilizadas en clasificación, se requiere una interpretación más precisa del significado de los conceptos inmersos en el historial, así como del contexto en el que intervienen, de cara a determinar inequívocamente el código correspondiente[20].

- Recuperación Automática de HCE similares a uno dado:

Resulta evidente que la información contenida en los historiales clínicos, no sólo del propio paciente sino de otros que hayan sufrido síntomas similares o hayan sido sometidos a los mismos tratamientos, puede ser de gran utilidad en el diagnóstico. Sin embargo, dada la elevada cantidad de historiales almacenados en clínicas y hospitales, la consulta de todos ellos resulta impensable. Es por ello que en los últimos años, y propiciado por la implantación obligatoria del Historial Clínico Electrónico en EEUU y su próxima implantación en toda la geografía de nuestro país, han proliferado las iniciativas, tanto académicas como empresariales, que abogan por construir sistemas que permitan la recuperación automática de historiales similares a uno dado.

Los principales problemas a los que se enfrentan este tipo de sistemas son dos. En primer lugar, la necesidad de disponer de una gran cantidad de historiales anonimizados para la experimentación y validación; y en segundo lugar, las particularidades del lenguaje utilizado en su redacción. Constantemente, y especialmente en las urgencias sanitarias, las presiones de tiempo hacen que el médico escriba utilizando elisiones, abreviaciones y omisiones gramaticales. Además, a menudo se utilizan distintos términos para designar un mismo concepto (sinonimia) o términos que pueden significar conceptos muy distintos dependiendo de su contexto (homonimia). Aunque, como se estudia en la siguiente sección, algunos de estos problemas pueden ser solventados utilizando los recursos lingüísticos adecuados, para otros aún no se ha encontrado una solución completamente satisfactoria.

La gran mayoría de los sistemas tradicionales de recuperación de historias clínicas basan sus decisiones en los términos identificados en el texto, y sólo algunos de ellos utilizan listas a medida o etiquetado manual para determinar cuáles de estos términos corresponden a síntomas, enfermedades u otras características relevantes[21]. Sin embargo, pocas explotan las posibilidades que ofrecen los tesauros y terminologías clínicas a la hora de realizar un análisis semántico del texto que, como se ha demostrado, presenta dos ventajas fundamentales. En primer lugar, evita el coste que supone la elaboración y constante actualización de la lista de síntomas. En segundo lugar, y más importante, permite obtener y precisar el significado de los términos así como capturar las posibles relaciones semánticas que se establecen entre los mismos (sinonimia, homónima, hiperonimia, etc.). De este modo, se consigue basar la recuperación en una representación semántica del texto, enriqueciendo significativamente el proceso de búsqueda, y solventando algunos de los problemas comentados.

- Generación automática de resúmenes:

En un contexto como el que se ha venido describiendo a lo largo de toda la exposición, resulta evidente que la generación automática de resúmenes (GAR) puede ser de gran utilidad. Durante el ejercicio de la medicina, disponer de resúmenes de los historiales de los pacientes puede ayudar a los profesionales a actuar con mayor celeridad en el tratamiento de urgencias sanitarias; mientras que, durante la formación y la investigación, los resúmenes pueden ser útiles para determinar si un documento resulta de interés, y si merece o no una lectura exhaustiva.

En la actualidad, pocos son los sistemas de administración clínica que incorporan entre sus funciones la generación de resúmenes de los documentos con los que tratan, y ello se debe fundamentalmente al estado preliminar en que se encuentran estos sistemas, que aún presentan muchas limitaciones y muchos problemas sin resolver. Por un lado, determinar qué información de un documento es relevante no es fácil, y además puede depender de las necesidades concretas de cada usuario. Más complicado aún resulta mantener la coherencia y la cohesión gramatical del resumen generado, que con frecuencia consiste en una sucesión de oraciones extraídas directamente del documento original.

Las investigaciones actuales en GAR en biomedicina generalmente se centran en resumir literatura científica. Una clasificación de alto nivel de estos sistemas es la que distingue entre aquellos que utilizan técnicas de extracción; es decir, generan resúmenes compuestos íntegramente por material del documento original, y aquellos que utilizan técnicas de abstracción; es decir, generan resúmenes que incluyen contenidos que no están explícitamente presentes en la entrada. Aunque típicamente los humanos realizan resúmenes mediante abstracción, la realidad es que la mayor parte de la investigación hoy día sigue siendo en extracción.

Los sistemas basados en extracción de oraciones realizan un análisis superficial del texto a nivel morfológico y sintáctico. Los primeros trabajos se limitaban a localizar segmentos clave en el documento original, utilizando criterios estadísticos, como la frecuencia de las palabras en el texto[22,23]; criterios posicionales, que tienen en cuenta la posición que ocupa cada oración en el documento[24]; y criterios lingüísticos, que evalúan la presencia de determinadas expresiones o palabras indicativas[23]. En los últimos años, y obedeciendo al mismo razonamiento subrayado para la recuperación de HCE, han cobrado relevancia los enfoques que utilizan terminologías y ontologías para identificar los conceptos del documento y sus relaciones semánticas como paso previo para delimitar la información relevante[25,26]. Aún lejos de conseguir los resultados deseados, estos enfoques presentan un mejor desempeño al trabajar en el nivel semántico en lugar de en el nivel léxico, permitiendo identificar relaciones semánticas como la sinonimia o la homonimia.

4. Particularidades del lenguaje médico y limitaciones en la detección de conceptos

El lenguaje específico utilizado en los textos médicos, ya se trate de literatura científica o de historiales de pacientes, presenta una serie de características que, como se ha subrayado en la sección anterior, pueden socavar el rendimiento de los sistemas de acceso a la información[27].

- En primer lugar, se han de tener en cuenta las ***negaciones significativas***, noción que hace referencia al hecho de que la mera presencia de un concepto en un texto no significa que éste sea importante; sino que puede referirse, por ejemplo, a la ausencia del mismo o a su ocurrencia en un pasado remoto. La importancia de su adecuado reconocimiento es especialmente evidente en los sistemas de recuperación automática de HCE. Supóngase que se desea recuperar historiales similares a uno en el que aparece el texto *"El paciente presenta un síndrome viral respiratorio prolongado. El examen físico no revela neumonía"*. Si no se detecta correctamente la negación que acompaña al concepto neumonía, el sistema recuperará un elevado porcentaje de informes de pacientes que hayan sufrido esta enfermedad. Son muchas las soluciones a este problema que se han propuesto en la literatura, la mayoría basadas en el uso de patrones para detectar la ausencia de síntomas y/o enfermedades.
- Un segundo problema que a menudo se plantea a la hora de delimitar los conceptos biomédicos presentes en un documento es el de los ***sinónimos*** (utilización de diferentes términos para designar un mismo concepto) y los ***homónimos*** (uso de términos con múltiples significados). Por ejemplo, los términos *infarto* y *paro cardíaco* comparten significado; mientras que el término *anestesia* puede hacer referencia tanto a la pérdida de la sensibilidad dolorosa como al procedimiento o fármaco utilizado para inducirla. En relación al fenómeno de la sinonimia, su solución es relativamente sencilla si se utiliza alguna ontología o terminología como UMLS, ya que en ellas ambos términos se presentan asociados a un mismo concepto. Por el contrario, la solución al problema de la homonimia es más compleja, ya que requiere el uso de algún algoritmo de desambiguación semántica que asigne a cada término

el significado pertinente en función del contexto en el que se halle inmerso.

- Otro problema al que se enfrentan las aplicaciones informáticas al tratar con textos médicos es la frecuente presencia de ***neologismos***, ***elisiones*** y ***abreviaciones***. Una ***elisión*** es una oración en la que se ha suprimido algún elemento del discurso, sin que ello conlleve consecuencias gramaticales (por ejemplo, en la oración "*la cuenta blanca fue de 1800*", la expresión *cuenta blanca* hace referencia al recuento de glóbulos blancos). Un ***neologismo*** es el uso de un nuevo vocablo o término que no se espera se encuentre en diccionarios o terminologías (por ejemplo, los términos *positividad* o *sobrecrecimiento*). Finalmente, una ***abreviación*** es una reducción de una palabra, mediante la supresión de determinadas letras o sílabas, que puede crear confusión a la hora de reconocer automáticamente los conceptos inmersos en el texto (por ejemplo, *adenoca.* se utiliza con frecuencia en lugar de *adenocarcinoma;* mientras que *AH* se utiliza habitualmente en lugar de *ácido hialurónico*). De nuevo la utilización de ontologías, en particular UMLS, es de gran utilidad a la hora de resolver estos fenómenos lingüísticos, ya que contiene numerosas entradas de elisiones, neologismos y abreviaciones que, además, se encuentran en continua actualización.

5. Conclusiones

La sobrecarga de información a la que se enfrentan los médicos en su quehacer diario es cada vez mayor. Si bien en sus inicios la digitalización prometía una simplificación de su trabajo, lo cierto es que en un entorno así la búsqueda de informaciones concretas y relevantes se hace difícil, y el exceso de información puede dificultar la toma de decisión y afectar a la calidad del trabajo realizado. Ahora bien, si la informática y las telecomunicaciones han contribuido a esta sobrecarga cognitiva, también pueden proporcionar las herramientas necesarias para aprovechar mejor todo este conocimiento a disposición de los médicos.

Los avances conseguidos en PLN e IA hacen cada vez más factible el desarrollo de sistemas inteligentes que faciliten el acceso a la información biomédica a distintos niveles y para muy diversas fuentes. Así, los sistemas de recuperación de información suministran mecanismos automáticos de búsqueda de documentos relevantes a los términos de una consulta, que posteriormente pueden ser clasificados en distintas categorías utilizando técnicas de clasificación automática. En una etapa posterior, los sistemas de generación automática de resúmenes pueden procesar estos mismos documentos recuperados para construir un resumen que sintetice la información significativa y permita así reducir el tiempo empleado en la adquisición del conocimiento. Finalmente, los sistemas de recuperación automática de HCE podrían simplificar la labor de los médicos en la búsqueda de evidencias en casos similares.

Sin embargo, y a pesar de los avances que prometen estas tecnologías, su nivel de implantación en hospitales y centros de investigación es todavía muy reducido; y algunas de ellas, como la GAR o la recuperación automática de HCE no han transcendido aun del ámbito académico.

Los motivos de esta escasa difusión son muy diversos, y van desde la deficiente formación de los profesionales en la selección, gestión y uso de estas tecnologías, hasta la falta de comprensión de las necesidades de los servicios de salud por parte de los encargados de desarrollarlas.

Precisamente uno de los principales obstáculos lo constituye la falta de estandarización de la documentación que se produce, tanto en relación a la estructuración de los contenidos como en lo que a uso de una terminología controlada se refiere. Si bien los expertos en informática médica han conseguido considerables progresos en los últimos años en el desarrollo de estándares para historias clínicas (CEN, HL7), nomenclaturas (SNOMED, Read Codes), clasificaciones (la Clasificación Internacional de Enfermedades de la OMS y la Clasificación Internacional de Problemas de Salud de WONCA) y tesauros o lenguajes controlados (MeSH), lo cierto es que aún queda mucho trabajo por hacer.

Agradecimientos

Esta investigación está financiada por el Ministerio de Ciencia e Innovación a través del programa FPU y por la CAM y el Fondo Social Europeo a través del programa IV PRICIT.

Bibliografía

1. Hersh W. Information Retrieval. A Health and Biomedical Perspective. 3rd Edition. Health Informatics Series. Springer; 2009.
2. Fauci A, Braunwald E, Kasper D, Hauser S, Longo D, Jameson. Harrison Principios de Medicina Interna. 17ª Edición. McGraw- Hill; 2009.
3. Cohen K, Hunter L. Getting Started in Text Mining. Olga Troyanskaya, Editor. Princeton University. 2008.
4. Bondenreider O, Mitchell J, McCray A. Biomedical Ontologies. In Proceedings of the 2003 Pacific Simposium on Biocomputing. World Scientific: 562-564. 2003.
5. SNOMED International. Disponible en: http://www.snomed.org/snomedct
6. NLM Unified Medical Language System (UMLS). Disponible en: http://www.nlm.nih.gov/research/umls
7. NLM Medical Subject Headings (MeSH). Disponible en: http://www.nlm.nih.gov/mesh/
8. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. Journal of the American Medical Informatics Association. 1994; 1: 447- 458.
9. Salton G, Wong A, Yang C.S. A Vector Space Model for Automatic Indexing. *Communications of the ACM*. 1975; 18(11): 613–620.
10. Kwangcheol S, Sang-Yong H. Improving Information Retrieval in Medline by Modulating MeSH Term Weights. Lecture Notes in Computer Science. Springer. 2004; 3136: 388-394.
11. Medina R.C. Semantic Information Retrieval: Representing and querying a heterogeneous biological documentary memory. E-gnosis. 2004; Vol. II.
12. Cohen A.M, Hersh W.R. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005; 6 (1): 57-71.
13. Lee K.P, Schotland M, Bacchetti P, Bero L.A. Association of Journal Quality Indicators with Methodological Quality of Clinical Research Articles. Journal of the American Medical Association. 2002; 287(21): 2805-2808.
14. Nelson S.J, Aronson A.R, Doszkocs T.E, Wilbur W.J, Bodenreider O. Automated Assignment of Medical Subject Headings. Proceedings of AMIA Annual Symposium .1999; 1127.
15. Darmoni S.J, Névéol A, Renard J.M, Gehanno J.F, Soualmia L.F, Dahamnal B, Chang H.F. A MEDLINE categorization algorithm. BMC Medical Informatics and Decision Making 2006.
16. Singhal A. Modern information retrieval: a brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001; 24: 35–43.
17. Manber U. and Wu S. GLIMPSE: a tool to search through entire file systems. Proceedings of the USENIX Conference. 1994; 23–32.
18. Dobrokhotov P. B, Goutte C, Veuthey A, Gaussier E. Combining NLP and probabilistic categorisation of document and term selection for Swiss-Prot medical annotation. Bioinformatics. 2003; 19 (1): 91–94.
19. Computational Medicine Center Medical Natural Language Processsing Challenge 2007 (CMC-NLP 2007). Disponible en: http://www.computationalmedicine.org/challenge.
20. Sotelsek-Margalef A, Villena-Román J. MIDAS: An Information-Extraction Approach to Medical Text Classification. Procesamiento del Lenguaje Natural. 2008; 41: 97-104.
21. Kwiatkowska M, Atkins S. Case Representation and Retrieval in the Diagnosis and Treatment of Obstructive Sleep Apnea: A Semiofuzzy Approach. Proceedings European Case Based Reasoning Conference, ECCBR 04. 2004.
22. Luhn H. P.The Automatic Creation of Literature Abstracts. IBM Journal of Research Development. 1958; 2(2): 159-165.
23. Edmundson H.P. New Methods in Automatic Extracting. Journal of the Association for Computing Machinery. 1969; 2(16): 264-285.Brandow, R., K. Mitze, y L. F. Rau. Automatic Condensation of Electronic Publications by Sentence Selection. Information Processing and Management. 1995; 5(31):675-685.
24. Plaza L, Díaz A, Gervás P. Uso de Grafos de Conceptos para la Generación Automática de Resúmenes en Biomedicina. Procesamiento del Lenguaje Natural. 2008; 41: 191-198.
25. Yoo I, Hu X, Song I.Y. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinformatics. 2007; 8(9).
26. Nadkarni P.M. Information Retrieval in Medicine: Overview and Applications. Journal of Postgraduate Medicine. 2000; 46(2): 122-166.