Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 189-192 recibido 02-04-2018 revisado 26-04-2018 aceptado 04-05-2018

Buscador Semántico Biomédico

Biomedical Semantic Information Retrieval

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Arturo Montejo-Ráez, Fernando Martínez-Santiago, Alberto Andreu-Marín, M. Teresa Martín-Valdivia, L. Alfonso Ureña López
SINAI Group - CEATIC - Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
{plubeda, mcdiaz, amontejo, dofer}@ujaen.es
{aandreu, maite, laurena}@ujaen.es

Resumen: El Buscador Semántico Biomédico propone una herramienta web de sencilla utilización para la identificación de terminología médica, la recuperación de literatura especializada y la exploración semántica del contenido gracias a la integración, con requisitos de tiempo de respuesta y alta disponibilidad, de ontologías médicas, técnicas de análisis de texto, reconocimiento de entidades y búsqueda de información sobre fuentes diversas externas. El resultado es una aplicación intuitiva y a la vez potente, que permite identificar la terminología médica sobre cualquier texto con un solo "click". Sobre ese reconocimiento, se permite filtrado de sentidos y subconceptos y la recuperación de información sobre recursos como SciELO, Google Scholar y Medline. Además, el sistema genera un grafo conceptual de manera automática, que permite relacionar semánticamente los términos que aparecen en el texto.
Palabras clave: reconocimiento automático de entidades, ontologías, terminología médica, informes médicos, recuperación de información, UMLS

Abstract: The Biomedical Semantic Information Retrieval system is an easy web solution to medical term identification, retrieval of specialized literature and semantic concept browsing thanks to the integration, with constraints in speed and high availability, of medical ontologies, text analysis, entity recognition and information retrieval from multiple sources. The result is an intuitive application, yet powerful, that performs term identification of medical concepts over any text with a simple click. Over identified terms the user is able conduct sub-concept selection to fine-tune the retrieval process over resources like SciELO, Google Scholar and Medline. Besides, the system generates a conceptual graph automatically which semantically relates all the terms found in the text.
Keywords: automatic entity recognition, term identification, ontologies, biomedical terminology, medical reports, information retrieval, UMLS

1 Introducción

Las herramientas para la comprensión de terminología especializada permiten, por un lado, ayudar a los iniciados en una mejor legibilidad de los textos y, para los expertos, acceder a nivel mayor de análisis cuando esas herramientas proveen de referencias adicionales sobre los términos detectados. Los sistema de detección de terminología médica han despertado siempre mucho interés (Krauthammer and Nenadic, 2004), dado el elevado número de recursos relacionados y la calidad de éstos, como las ontologías especializadas UMLS (Bodenreider, 2004) y MeSH (Díaz-Galiano et al., 2007). La identificación de términos no es sólo interesante sobre textos especializados, sino que puede resultar muy útil también sobre los textos escritos por pacientes (MacLean and Heer, 2013). En ambos casos, los sistemas de identificación de términos se convierten en la clave para acceder a documentación bibliográfica y literatura relacionada, pues estos términos y sus relaciones, toda vez identificados, permiten conectar el conocimiento entre

ISSN 1135-5948. DOI 10.26342/2018-61-29

distintas fuentes.

La plataforma propuesta es lo suficientemente flexible como para que puedan beneficiarse de ella tanto profesionales de medicina como usuarios no especializados (por ejemplo, pacientes).

La búsqueda semántica busca mejorar la precisión de la búsqueda al comprender la intención del usuario y el significado contextual de los términos tal y como aparecen en el espacio de búsqueda, ya sea en la Web o dentro de un sistema cerrado y así generar al lector los resultados más relevantes.

2 *Desarrollo del prototipo*

El prototipo es un sistema de recuperación de información sobre textos biomédicos. Identifica terminología especializada automáticamente y la utiliza en un proceso de meta-búsqueda sobre varias bases documentales, enriqueciendo semánticamente los resultados para llegar a obtener mayor precisión. La figura 1 muestra un esquema de su funcionamiento.

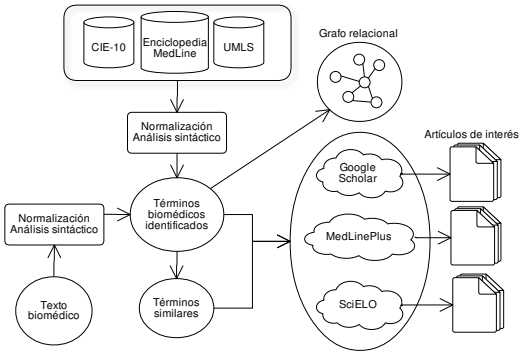

Figura 1: Diagrama de funcionamiento del prototipo

La aplicación recibe como entrada un texto, bien mediante un formulario, bien mediante una extensión para el navegador Google Chrome que permite lanzar el procesamiento desde una simple selección de texto sobre cualquier página web. Se reconocen entidades tales como enfermedades, síntomas y tratamientos, entre otros. Además, el prototipo permite incluir bases de conocimientos de forma rápida y sencilla. El sistema se ha construido sobre tecnologías como Elasticsearch, Google Scholar y el *webservice* de Medline, que pasamos a detallar a continuación.

Elasticsearch[1] es una potente herramienta que nos permite indexar una gran volumen de datos y posteriormente hacer consultas sobre ellos soportando entre otras muchas cosas búsquedas aproximadas, facetas y resaltado.

Google Scholar[2] es un buscador de Google enfocado en el mundo académico que se especializa en literatura científico-académica en una variedad de disciplinas y formatos de publicación. El índice de Google Scholar incluye la mayoría de las revistas y libros académicos en línea revisados por pares, documentos de conferencias, tesis y disertaciones, preimpresiones, resúmene, informes técnicos y otra literatura académica, incluyendo opiniones judiciales y patentes.

MedlinePlus[3] es un servicio de información en línea provisto por la Biblioteca Nacional de Medicina de los Estados Unidos. MedlinePlus contienen enlaces a portales de Internet con información de alrededor de más de 1.000 temas de salud, además, estos temas de salud incluyen enlaces a noticias actualizadas diariamente. Este recurso es de interés para usuarios menos expertos ya que ofrece un vocabulario más informal y familiar para el lector.

Gracias al desarrollo modular del prototipo, es posible independizar diversos aspectos en diferentes servicios incluidos en sistemas potentes que permitan mayor rendimiento sobre bases de conocimiento más amplias. La figura 2 muestra el aspecto final de este prototipo.

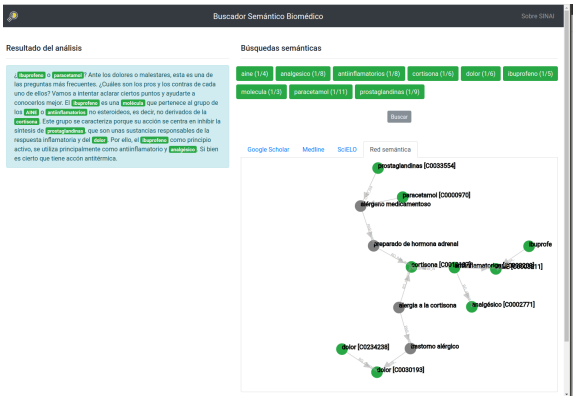

Figura 2: Captura de pantalla del prototipo

2.1 Reconocimiento de entidades

Existen herramientas que reconocen entidades médicas en inglés, como MetaMAP (Aronson, 2001) o cTakes (Savova et al., 2010), también para el español como la versión en español de

[1] https://www.elastic.co/

[2] https://scholar.google.es/

[3] https://medlineplus.gov/

MMTx (Carrero, Cortizo, and Gómez, 2008) y Freeling-Med (Oronoz et al., 2013), si bien el uso de estas tecnologías en un herramienta con requisitos de respuesta se tornó poco conveniente. Por lo tanto, se ha diseñado una herramienta propia para la detección de términos biomédicos dentro del texto. Los recursos y algoritmos que utiliza el reconocedor son:

- **Bases de conocimiento**: UMLS (*Unified Medical Language System*), es un compendio de diversos vocabularios y estándares, tanto de salud como biomédicos, para permitir la interoperabilidad entre sistemas informáticos. UMLS permite, entre otros usos, vincular información de salud, términos médicos, nombres de medicamentos y códigos de facturación a través de diferentes sistemas informáticos. Un ejemplo de uso es la vinculación de términos y códigos entre médico, farmacia y compañía de seguros a través de la enciclopedia de Medline y CIE-10.
- **Procesamiento del lenguaje natural**: Para la detección de entidades se ha llevado a cabo una normalización del texto, tanto en los diccionarios utilizados como en el texto introducido por el usuario. La herramienta utilizada en este caso es la biblioteca NLTK (*Natural Language Toolkit*) desarrollada en el lenguaje de programación Python. Además, para obtener una mayor precisión a la hora de identificar terminología (lematización, desambiguación, palabras compuetas), se utiliza el analizador sintáctico incluido en la herramienta CoreNLP desarrollada por la Universidad de Stanford para el español (Manning et al., 2014).
- **Referencias enlazadas**: Dentro del dominio clínico, se hace notar la importancia acerca del intercambio de información debido a los nuevos requisitos de los servicios sanitarios. Para poder seguir prestando servicios como cambios demográficos, movilidad o equidad en el acceso a información de forma efectiva y eficiente es necesario el uso de catálogos estandarizados internacionalmente que unifiquen los datos empleados en las distintas instituciones (Hammond and Cimino, 2006).

  El prototipo muestra los conceptos médicos detectados junto con su código identificador según el diccionario del que se haya extraído la información, este código, además incluye una referencia a una web. Así por ejemplo, el término *'cólera'* tiene asociado el código *C0008354*, *A00* y *000303* en UMLS, CIE-10 y la enciclopedia de MedLine respectivamente.

2.2 Búsqueda semántica

La ontología UMLS proporciona para cada concepto detectado otros términos con significado similar. Todos los términos similares contienen el mismo CUI (*Concept Unique Identifier*), por ejemplo, para el código C0004057 se obtienen los conceptos 'Aspirina', 'Ácido acetilsalicílico', 'AAS' o 'aspirina como antiplaquetario', entre otros. Con ello, el usuario puede filtrar su consulta con los términos que él considere apropiados para su búsqueda.

Partiendo de esta terminología seleccionada, se lanzan consultas sobre varias bases documentales como, Google Scholar, Medline o SciELO. Los términos origen de la consulta pueden ser en todo momento modificados mediante la selección de terminología más específica o sinónimos gracias a la posibilidad de personalizar cada término a partir del concepto detectado, como se ha explicado más arriba. Inmediatamente después es posible relanzar las consultas a las distintas fuentes, obteniendo así nuevos resultados.

2.3 Grafo relacional

Gracias a que la información que manejamos es muy rica en contenido semántico, una opción añadida al prototipo BSB es modelar dicha información como una red semántica. En UMLS, los términos sinónimos se agrupan para formar un mismo concepto y los conceptos se vinculan unos a otros por medio de varios tipos de relaciones, lo que da como resultado un gráfico enriquecido, etiquetado y dirigido.

Para la generación, se hace uso del algoritmo de Dijkstra también llamado algoritmo de caminos mínimos sobre grafos. Es un algoritmo para la determinación del camino más corto desde un vértice inicial al resto de vértices en un grafo. El vértice origen es el llamado nodo central o centroide, este centroide se elige según la frecuencia de aparición en el texto.

El grafo es interactivo, por lo que el usuario puede elegir cualquier otro nodo como centroide lo cual disparará la regeneración del grafo de forma distinta a la anterior. De esta manera es posible explorar la ontología UMLS

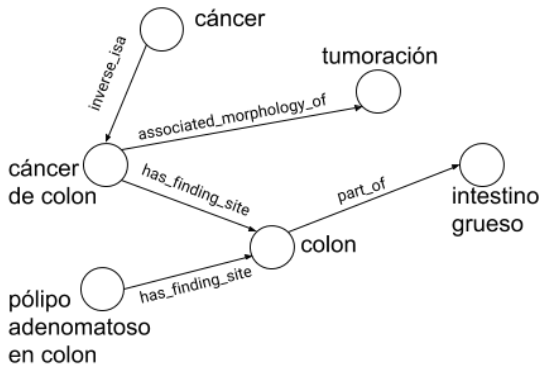

Figura 3: Grafo semántico sobre seis conceptos identificados a partir de un texto

de forma visual e interactiva.

Dentro del metatesauro UMLS, las relaciones simbólicas pueden ser jerárquicas (por ejemplo, '*is a kind of*', '*isa*', '*part of*') o asociativa (por ejemplo, '*location of*', '*caused by*'). En la Figura 3 podemos observar cómo una búsqueda simple de cáncer de colon nos lleva a obtener información relevante acerca de los nodos más cercanos a través de sus relaciones como *has finding site* o *part of*.

3 Conclusiones

El objetivo de la aplicación desarrollada es ofrecer una interfaz sencilla que permita, tanto al experto en medicina como a un usuario no experto, acceso a conocimiento adicional y potencialmente útil a partir de un texto biomédico. El sistema presenta el resultado del análisis (detección de entidades), los resultados de búsqueda en varias fuentes documentales y la red semántica de conceptos identificados, todo esto con respuesta en tiempo real. El sistema es interactivo, pues las búsquedas pueden refinarse a partir de los términos asociados a los conceptos identificados y sobre el grafo es posible explorar las relaciones entre términos de una forma visual y modificar los términos centrales para la construcción del mismo.

Agradecimientos

Este trabajo está parcialmente subvencionado por el proyecto REDES (TIN2015-65136-C2-1-R) del MICINN del Gobierno de España.

Bibliografía

Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of AMIA*, page 17. American Medical Informatics Association.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Carrero, F., J. C. Cortizo, y J. M. Gómez. 2008. Building a spanish mmtx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer.

Díaz-Galiano, M. C., M. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, y L. Urena-López. 2007. Integrating mesh ontology to improve medical information retrieval. In *Workshop of the CLEF*, pages 601–606. Springer.

Hammond, W. E. y J. J. Cimino. 2006. Standards in biomedical informatics. In *Biomedical Informatics*. Springer, pages 265–311.

Krauthammer, M. y G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526.

MacLean, D. L. y J. Heer. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of AMIA*, 20(6):1120–1127.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, y D. McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of ACL: system demonstrations*, pages 55–60.

Oronoz, M., A. Casillas, K. Gojenola, y A. Perez. 2013. Automatic annotation of medical records in spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536–543. Springer.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, y C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of AMIA*, 17(5):507–513.