Sistemas de recuperación de información adaptados al dominio biomédico

Por **Mónica Marrero**, **Sonia Sánchez-Cuadrado**, **Julián Urbano**, **Jorge Morato** y **José-Antonio Moreiro**

Resumen: *La terminología usada en biomedicina tiene rasgos léxicos que han requerido la elaboración de recursos terminológicos y sistemas de recuperación de información con funciones específicas. Las principales características son las elevadas tasas de sinonimia y homonimia, debidas a fenómenos como la proliferación de siglas polisémicas y su interacción con el lenguaje común. Los sistemas de recuperación de información en el dominio biomédico utilizan técnicas orientadas al tratamiento de estas peculiaridades léxicas. Se revisan algunas de estas técnicas, como la aplicación de Procesamiento del Lenguaje Natural (BioNLP), la incorporación de recursos léxico-semánticos, y la aplicación de Reconocimiento de Entidades (BioNER). Se presentan los métodos de evaluación adoptados para comprobar la adecuación de estas técnicas en la recuperación de recursos biomédicos.*

Palabras clave: *Biomedicina, BioNER, BioNLP, Text-mining, Recuperación de información, Proceso del lenguaje natural, NLP.*

Title: Information retrieval systems adapted to the biomedical domain

Abstract: *The terminology used in biomedicine has lexical characteristics that have required the elaboration of terminological resources and information retrieval systems with specific functionalities. The main characteristics are the high rates of synonymy and homonymy, due to phenomena such as the proliferation of polysemic acronyms and their interaction with common language. Information retrieval systems in the biomedical domain use techniques oriented to the treatment of these lexical peculiarities. In this paper we review some of these techniques, such as the application of Natural Language Processing (BioNLP), the incorporation of lexical-semantic resources, and the application of Named Entity Recognition (BioNER). Finally, we present the evaluation methods adopted to assess the suitability of these techniques for retrieving biomedical resources.*

Keywords: *Biomedicine, BioNER, BioNLP, Text-mining, Information retrieval, Natural Language Processing, NLP.*

Marrero, Mónica; **Sánchez-Cuadrado, Sonia**; **Urbano, Julián**; **Morato, Jorge**; **Moreiro, José-Antonio**. "Sistemas de recuperación de información adaptados al dominio biomédico". *El profesional de la información*, 2010, mayo-junio, v. 19, n. 3, pp. 246-254.

DOI: 10.3145/epi.2010.may.04

Jorge Morato *es doctor en documentación y trabaja en el Departamento de Informática de la Universidad Carlos III de Madrid. Sus principales publicaciones están relacionadas con la construcción automática de tesauros y ontologías. Sus líneas de investigación se centran principalmente en proyectos sobre "semantic metadata interoperability and search".*

José-Antonio Moreiro *es catedrático de Biblioteconomía y Documentación de la Univ. Carlos III de Madrid (UC3M). Es autor de trabajos sobre técnicas de análisis de contenido documental y cuestiones conceptuales de la documentación. Ha sido director del Depto. de Biblioteconomía y Documentación y decano de la Fac. de Humanidades, Comunicación y Documentación de la UC3M.*

Julián Urbano *es ingeniero en informática y ha realizado un Máster en Ciencia y Tecnología Informática. Tras una beca como ayudante de investigación en Virginia Tech, actualmente es profesor ayudante del Departamento de Informática de la Universidad Carlos III de Madrid, donde cursa el doctorado. Su principal área de investigación es la evaluación en recuperación de información.*

Sonia Sánchez-Cuadrado *es doctora en documentación y trabaja en el Departamento de Informática de la Universidad Carlos III de Madrid (UC3M). Su actividad investigadora está enfocada a la extracción de información, reconocimiento de patrones, procesamiento del lenguaje natural, sistemas de organización del conocimiento y recuperación de información.*

Mónica Marrero *es ingeniera técnica en informática y licenciada en documentación. Ha realizado el Máster en ciencia y tecnología informática en la Universidad Carlos III de Madrid y actualmente trabaja como profesora ayudante en el Departamento de Informática de la misma universidad. Su investigación se centra en la extracción y recuperación de información.*

Introducción y características del dominio biomédico

El incremento de documentación y la urgencia para localizar respuestas relevantes convierten a la literatura biomédica en área de interés para la aplicación de técnicas de tratamiento textual en el campo de la recuperación de información (IR, *information retrieval*). Los sistemas de IR en biomedicina (tabla 1) utilizan voca-

Artículo recibido el 15-12-09
Aceptación definitiva: 08-02-09

Motores de recuperación	Colección	Características técnicas principales
Novoseek *http://www.novoseek.com*	*Medline*, *US Grants*, otros	Basadas en diccionarios elaborados automáticamente
PubFocus *http://www.pubfocus.com*	*PubMed*, *Medline*	Basadas en diccionarios (*NCI Thesaurus* y *Mouse genome database*). Utiliza el índice *Journal citation reports* en la ordenación de resultados.
BioMedSearch *http://www.biomedsearch.com*	*PubMed*, *Medline*	*Clustering*
XplorMed *http://www.ogic.ca/projects/xplormed/info*	*Medline*	NLP (usan *TreeTagger* para la eliminación de palabras vacías, la desambiguación y normalización con *stemming*)
Path Binder H (prototype) *http://pathbinderh.plantgenomics.iastate.edu*	*PubMed*, *Medline* (parcialmente)	Basadas en diccionarios (*Gene & plant ontology*, *Enzyme nomenclature* y *MeSH*). Incluye filtrado por taxonomía.
Textpresso *http://www.textpresso.org*	*C. Elegans*. Permite añadir nuevos artículos	Basadas en diccionarios de términos y procesos elaborados automáticamente a partir del corpus *Caenorhabditis Elegans*

Tabla 1. Ejemplos de motores de recuperación disponibles para información biomédica

bularios controlados (*MeSH*, *Inspec thesaurus*, *Gene & plant ontology*) para mejorar las búsquedas, siguiendo las propuestas de **Salton** en los años 60 para los sistemas de IR de propósito general.

Las técnicas de procesamiento de lenguaje natural (NLP, *natural language processing*), la extracción de información (IE, *information extraction*) y la minería de datos se han convertido en procesos indispensables para tratar, identificar e inferir información entre la enorme cantidad de datos disponibles. Este conjunto de técnicas, referidas habitualmente como minería de textos (*text mining*), son utilizadas en los sistemas de IR y adoptan ciertas peculiaridades al ser aplicadas en el área biomédica por las características del dominio y su terminología.

"La normalización de la terminología logra consistencia en el etiquetado y favorece el uso efectivo de herramientas de anotación semántica"

La terminología biomédica carece de patrones regulares que faciliten la identificación automática de términos, a pesar de que un tercio de sus ocurrencias son variantes, ya sean ortográficas o por permutación, inserción o eliminación (e. g. proteínas Foxp2 y Foxp3) (**Jacquemin**, 2001). Además, la investigación en biomedicina se caracteriza por estar dividida en subáreas de conocimiento especializadas y conceptualizadas desde diferentes puntos de vista, lo que limita las conexiones entre trabajos (**Weeber** et al., 2000) y la recuperación cruzada de información. La aparición constante de nuevos términos para un mismo concepto en las diversas áreas y los diferentes registros, así como la proliferación de acrónimos, generan una elevada polisemia y sinonimia léxica, lo que dificulta notablemente la recuperación de información.

Un concepto de este dominio puede tener seis o siete sinónimos porque surgen en áreas diferentes, bien por cuestiones comerciales, por falta de consenso entre los expertos o por evolución u obsolescencia científica (tabla 2).

1979-1982	*Immunologic deficiency syndrome*
1984-1986	*Human T-cell leukemia virus/HTLV/LAV*
1986-1992	*HIV*
1992-	*HIV-1/HIV-2*

Tabla 2. Descriptores de AIDS en MeSH desde su aparición

Por ejemplo, un producto farmacéutico como el paracetamol se conoce también con las denominaciones de DCI o acetaminofén. Su sinónimo en la nomenclatura *Iupac* (*International Union of Pure and Applied Chemistry*) es N-(4-hidroxifenil)etanamida, que es equivalente a la fórmula química $C_8H_9NO_2$, y que se representa con el código NO2 BE01 de la *ATC* (*Anatomical, therapeutic, chemical classification system* de la *OMS*). Además, su nombre comercial varía de un país a otro: en Estados Unidos es conocido con el nombre de *Tylenol* o *Datril*, en Inglaterra como *Tylex CD* o *Panadeine*, en España *Panadol*, *Termalgin*, *Efferalgan*, *Gelocatil* o *Apiretal*, y en Méjico *Tempra*.

Los acrónimos aumentan el número de sinónimos en la literatura científica, y además se estima que más del 80% son ambiguos (**Liu**; **Johnson**; **Friedman**, 2002). De hecho, cada cinco artículos en biomedicina

surge una nueva sigla, que llega a coincidir con un gran número de siglas preexistentes (**Spasic**; **Ananiadou**, 2005). La polisemia afecta tanto a nombres comunes (e. g. el término inglés "cold" puede significar "chronic obstructive lung disease", algo "frío" o "resfriado") como a terminología especializada de diferentes áreas de la biomedicina. Por ejemplo, NF2 es el nombre de un gen y de una proteína y una enfermedad relacionada con el mismo (**Bodenreider**, 2006). NFKB2 hace referencia a proteínas pertenecientes a especies diferentes: la de los humanos y la de los pollos.

Recursos para la representación de conocimiento en biomedicina

Para afrontar los problemas terminológicos se elaboran sistemas de organización del conocimiento tales como diccionarios, glosarios, clasificaciones, tesauros y ontologías (tabla 3). Sin embargo, esto no siempre simplifica los procesos de indización y recuperación de información. Por ejemplo, el vocabulario controlado *Unified medical language system* (*UMLS*), usado con la base de datos bibliográfica *Medline* proporciona expansión de consultas con términos relacionados. La expansión de la consulta con un vocabulario controlado mejora la efectividad, en especial con la expansión por términos sinónimos (**Hersh** et al., 2000), pero fenómenos como la polijerarquía pueden disminuir la precisión de los resultados. Por ejemplo, en *MeSH* el virus de la inmunodeficiencia humana puede encontrarse bajo los siguientes árboles jerárquicos: *RNA virus infections* [C02.782], *Sexually transmitted diseases* [C02.800], *Slow virus diseases* [C02.839] o *Immune system diseases* [C20].

http://www.ncbi.nlm.nih.gov/pubmed

"El incremento de documentación y la urgencia para localizar respuestas relevantes convierten a la literatura biomédica en campo de interés para la aplicación de técnicas avanzadas de IR"

La complejidad de estos recursos es mayor que en otras áreas por el propio conocimiento que debe representarse (por ejemplo, procesos celulares, interacciones entre proteínas, estructuras proteicas, etc.) y la diversidad de áreas a las que pertenece al mismo tiempo. Como consecuencia, los recursos de información en biomedicina resultan difíciles de construir, usar y mantener. A pesar de ello, al menos en biología parece que el entusiasmo por las ontologías ha venido acompañado de una general falta de conocimiento acerca de lo que son y de cómo se usan (**Soldatova**; **King**, 2005). La iniciativa *Open biomedical ontologies* (*OBO*), iniciada en 2001, trabaja para reconducir esta tendencia en el campo de la biomedicina, ofreciendo reglas para la construcción de ontologías en subáreas como anatomía, genómica, proteómica, metabolómica, fenotipos, etc.

http://obofoundry.org

Otras iniciativas se centran en la construcción de ontologías específicas del dominio de alto nivel, como *OBR* (*Ontology of biomedical reality*), utilizada para integrar ontologías de anatomía, fisiología y patología (**Rosse** et al., 2005). Hay múltiples ejemplos del uso de este tipo de ontologías, como la *Molecular biology ontology* (*MBO*) (**Schulze-Kremer**, 1997) o *BFO* (*Basic formal ontology*) y *RO* (*Relation ontology*), que dentro del proyecto *OBO* dan soporte a la construcción de otras ontologías. Por ejemplo, *BioTop* y *ChemTop* (**Stenzhorn** et al., 2008) son dos ontologías de alto nivel para biología y química basadas en ellas. *BFO* utiliza a su vez dos conocidas ontologías de alto nivel no específicas del dominio: *Dolce* y *SUMO*. *Dolce* es también la base para la ontología biomédica *Simple bio upper ontology* (*SBUO*) (**Rector**; **Stevens**; **Rogers**, 2006) y *SUMO* es la base de la ontología de alto nivel elaborada en el proyecto *BioCaster* para la vigilancia de enfermedades infecciosas en los países asiáticos (**Collier** et al., 2007), que integra diferentes fuentes de conocimiento en varios idiomas. Hay además meta-ontologías para el dominio biomédico, como es el caso de *Bio-Zen*, que unifica diferentes esquemas de representación: *Dolce*, *Simple knowledge organisation system* (*SKOS*), *Semantically interlinked open communities* (*SIOC*), *Friend of a friend* (*FOAF*), *Dublin core* y *Creative commons* (**Samwald**; **Adlassnig**, 2008).

En línea con los objetivos de la web semántica, en los últimos años se han producido avances con trabajos como *Bio2Rdf* y *Linking Open Drug Data Project*. Estos proyectos ofrecen la posibilidad de integrar esquemas sobre genes, proteínas, medicamentos y ensayos clínicos, tanto sobre otros esquemas específicos de biomedicina como sobre esquemas genéricos (por ejemplo, *Dbpedia*).

http://bio2rdf.org
http://esw.w3.org/topic/HCLSIG/LODD
http://dbpedia.org

Todas estas iniciativas inciden directamente en los resultados de los sistemas de IR y de IE. Dada la complejidad de la terminología biomédica, la aplicación de técnicas basadas en diccionarios en lugar de en sistemas de representación más sofisticados como las ontologías, podría ser una de las razones por las que hasta

Recurso	Dominio	Contenido	Caracterísicas principales
UMLS (*Unified medical language system*) *http://www.nlm.nih.gov/research/umls*	Biomédicos y sanitarios	> 1.000.000 términos	Utiliza una ontología de alto nivel para integrar diferentes recursos, entre ellos *Snomed*, *MeSH* y *GO*
Snomed-CT (*Systematized nomenclature of medicine clinical terms*) *http://snob.eggbird.eu*	Historiales clínicos	400.000 términos	En inglés y español es gratuito, pero no en alemán. Implementado en lenguaje *OWL.*
GO (*Gene ontology*) *http://www.geneontology.org*	Términos genéticos, agrupados por funciones, procesos y localizaciones	> 9.000 términos	El 50% de los términos pueden ser mapeados en *MeSH* y *Snomed* (**McCray**, 2002)
MeSH (*Medical subject headings*) *http://www.nlm.nih.gov/mesh*	Biomedicina, incluyendo enfermería, veterinaria y sistemas sanitarios	22.995 descriptores	Utilizado para indexación automática y manual de *Medline*

Tabla 3. Ejemplos de sistemas de organización del conocimiento disponibles en el área de biomedicina

ahora se han obtenido peores resultados en biomedicina respecto a otros dominios (**Spasic** et al., 2005).

"El procesamiento de lenguaje natural (NLP), la extracción de información (IE) y la minería de datos (DM) adoptan ciertas peculiaridades al ser aplicadas en biomedicina"

Técnicas aplicadas a los sistemas de IR biomédica

Los sistemas de recuperación de información aplican habitualmente tareas de NLP, como la descomposición del texto en palabras o términos, el etiquetado gramatical, la normalización y, aunque en menor medida, la desambiguación léxica o la resolución de correferencias. La descomposición del texto para identificar los términos debe ser tratada de modo diferente en el dominio biomédico, ya que no puede ser resuelta directamente en base a los espacios en blanco y los signos de puntuación entre palabras (por ejemplo, [3H]R1881 es un único término). Algunos trabajos señalan que los etiquetadores gramaticales adaptados a este dominio mejoran la efectividad (**Zhou** et al., 2004; **Clegg**; **Sheperd**, 2005), lo que justifica la adaptación de las herramientas de NLP para biomedicina. Este es el caso del etiquetador gramatical y semántico *Genia*, que ha sido entrenado no sólo sobre corpus periodísticos (*Wall Street journal corpus*) sino también biomédicos (*Genia corpus* y *PennBioIE*), permitiéndole trabajar mejor con distintos tipos de documentos biomédicos (**Tsuruoka**; **Tsujii**, 2004).

Otra técnica aplicada es la extracción de información, que ha recibido una especial atención en biomedicina (*BioNER*) durante la última década, principalmente por los nombres de genes y productos genéticos debido al macroproyecto *Genoma humano*. *BioNER* se aplica al reconocimiento de ADN, ARN, línea celular, tipo celular, mutaciones, propiedades de las estructuras proteicas, etc.

Los sistemas de *BioNER* (tabla 4) han seguido una evolución similar a los de dominio general. Han pasado de utilizar técnicas basadas en reglas elaboradas de forma manual, a sistemas basados en aprendizaje supervisado a partir de corpus anotados. Estos últimos, conocidos habitualmente como clasificadores, se basan en la creación automática de reglas a partir de ejemplos anotados en un corpus. En *BioNER* se utiliza sobre todo este tipo de aproximación, aunque el apoyo de recursos léxicos es más acentuado. El uso de este tipo de recursos proporciona buenas tasas de precisión, pero no de exhaustividad, debido a la constante incorporación de nuevos términos.

"Un concepto en biomedicina puede tener seis o siete sinónimos, y cada cinco artículos surge una nueva sigla que coincide con un gran número de siglas preexistentes"

Como contrapartida, las técnicas semi-supervisadas son menos frecuentes en *BioNER*, aunque hay trabajos que utilizan *bootstrapping* y aprendizaje activo. Esta técnica se aplica en dominios generales desde los años 90 y se basa en la utilización de unos pocos ejemplos iniciales como semillas. A partir de ellos se realiza un aprendizaje progresivo de patrones léxico-sintácticos, basados en las características de las entidades a capturar y su contexto. En el aprendizaje activo, por el contrario, es el propio sistema el que proporciona al usuario los datos con mayor incertidumbre para que los corrija o etiquete, de modo que puedan ser utilizados como nue-

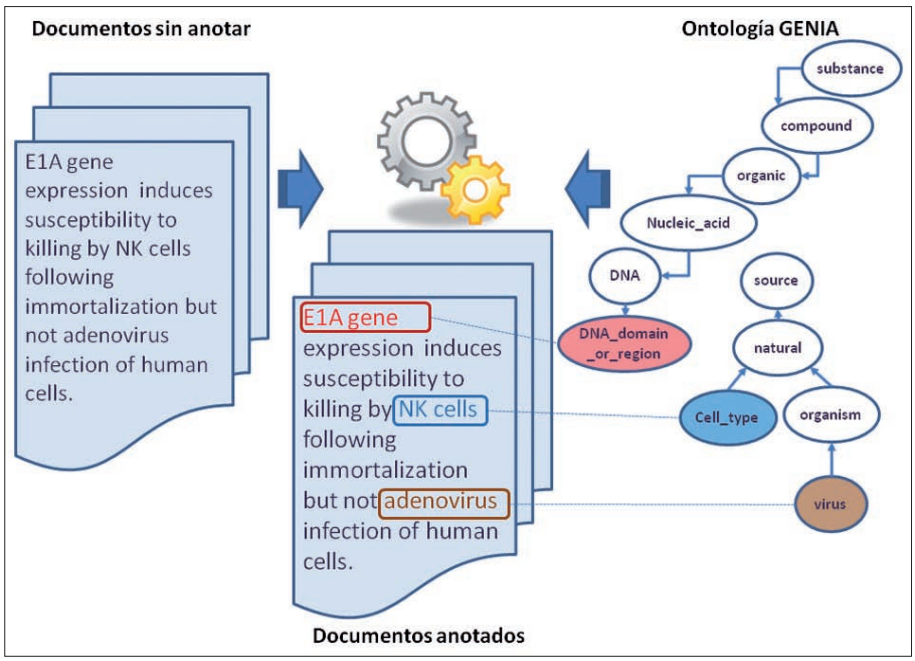

Figura 1. Ejemplo de anotación de un documento con la Ontología Genia

vas semillas. La aplicación de *bootstrapping* en *BioNER* es más reciente, y es habitual que tome como semillas corpus etiquetados completos a partir de diccionarios u ontologías (**Morgan** et al., 2004; **Rong** et al., 2009). La carencia de patrones en los términos y la existencia de contextos poco discriminantes por la proximidad semántica de algunas entidades (e. g. los genes se componen de proteínas, lo que hace que compartan procesos) dificultan el aprendizaje progresivo de patrones basado en un reducido número de ejemplos anotados. Por ello, el uso de recursos etiquetados más amplios se convierte en una parte esencial del proceso de aprendizaje.

La aplicación de estos métodos a la Web es poco frecuente, aunque el número de apariciones de ciertos patrones sintácticos se ha utilizado en aprendizaje supervisado (**Dingare** et al., 2004). En cualquier caso, las tendencias en minería de datos están ligadas a la evolución de la misma Web (**Baeza-Yates**, 2009). La creciente anotación semántica de la Web con métodos de la web semántica y de la web 2.0 sin duda contribuye a la mejora de los resultados de las técnicas de IE.

Otra de las técnicas de extracción de información frecuentemente aplicadas a los sistemas de IR es la detección de relaciones. Tareas como la detección de propiedades funcionales de genes o interacciones de proteínas están adquiriendo especial relevancia. En estas tareas se suman los problemas de *BioNER* a los múltiples tipos de relaciones diferentes que podemos encontrar. Por este motivo, el apoyo de sistemas de organización del conocimiento es también aquí más necesario que en otras áreas.

Evaluación de la IR en biomedicina

En las conferencias *TREC* (*Text retrieval conference*) se utilizaban colecciones de prueba del dominio médico para la evaluación de los sistemas de IR, pero hasta el año 2000 no se creó un *track* específico para biomedicina. Se midió la capacidad de distintos sistemas para clasificar los documentos de *OhsuMed* (subconjunto de *Medline*, con siglas procedentes de *Oregon Health Sciences University*) con las categorías de *MeSH*.

En 2003 surgió el *Genomics track* para la recuperación de documentos relevantes relacionados con genes. En el año 2004 este *track* se centró en la anotación de genes y proteínas, y se intentó emular el proceso manual que los anotadores del *Mouse genome informatics* realizan para anotar los genes con *GO* (**Hersh**, 2004). La última edición de este *track* tuvo lugar en 2007, y en él se requería responder a preguntas con entidades cuyo tipo era definido en la propia pregunta (e. g. *What [drugs] have been tested in mouse models of Alzheimer's disease?*). Otro foro importante en esta área es *BioCreative*, surgido en 2004 y centrado en el reconocimiento de genes e interacciones entre proteínas. Hay además foros para otro tipo de tareas, como *ImageClef* para la recuperación de imágenes médicas.

http://www.informatics.jax.org/
http://www.biocreative.org
http://ir.shef.ac.uk/imageclef

Bootstrapping

Literalmente significa correa (*strap*) de arranque (*to boot*), abreviado a veces en informática a *booting* (botar –un sistema operativo-). Este concepto se usa en diferentes áreas (electrónica, negocios, finanzas), y siempre significa poner en marcha algo con pocos medios, partiendo de un inicio muy simple.

"La evaluación de los sistemas de recuperación y extracción de información en biomedicina deben orientarse al usuario, con métricas y métodos capaces de medir su valor en escenarios reales"

Herramienta	Entidades	Características principales
Abner *http://pages.cs.wisc.edu/~bsettles/abner/*	Proteinas, ADN, ARN, línea y tipo celular	Aprendizaje supervisado (sobre Nlpba y *BioCreative*). Entrenable.
AbGene *http://ftp.ncbi.nlm.nih.gov/pub/tanabe/AbGene*	Genes, proteinas	Basada en reglas extraídas estadísticamente (sobre resúmenes de *Medline*)
PIE *http://pie.snu.ac.kr*	Proteinas, interacciones entre proteinas	NLP y uso de técnicas basadas en diccionarios y aprendizaje automático
Biorat *http://bioinf.cs.ucl.ac.uk/?id=754*	Proteinas, interacciones entre proteinas	NLP y uso de técnicas basadas en diccionarios y expresiones regulares. Utiliza el framework de *IE GATE. http://gate.ac.uk*
Lingpipe *http://alias-i.com/lingpipe/web/download.html*	Genes, proteinas y otros	Herramienta general de IE basada en aprendizaje supervisado (sobre *Genia* y *MedPost* para biomedicina). Entrenable.

Tabla 4. Ejemplo de herramientas BioNER disponibles

Las tasas de precisión alcanzadas para tareas de recuperación y extracción de información biomédica se encuentran entre un 70-90%, mientras que las de exhaustividad rondan el 70%. Estos resultados suponen un 15% menos respecto a las tasas alcanzadas en otros dominios como el periodístico (**Ananiadou**; **McNaught**, 2006). Sin embargo, las alcanzadas para NER (*Named entity recognition*) en el dominio periodístico no son superiores. En ocasiones se ha considerado incluso superada, con un 90% de acierto (**Cunningham**, 2005).

En estas tareas se ha demostrado que el cambio de las fuentes utilizadas, incluso cambiando únicamente el género documental y no el dominio, ocasiona una importante pérdida de efectividad (20-40%) (**Poibeau**; **Kosseim**, 2001). En el dominio biomédico se ha comprobado que entrenar con un corpus anotado y evaluar

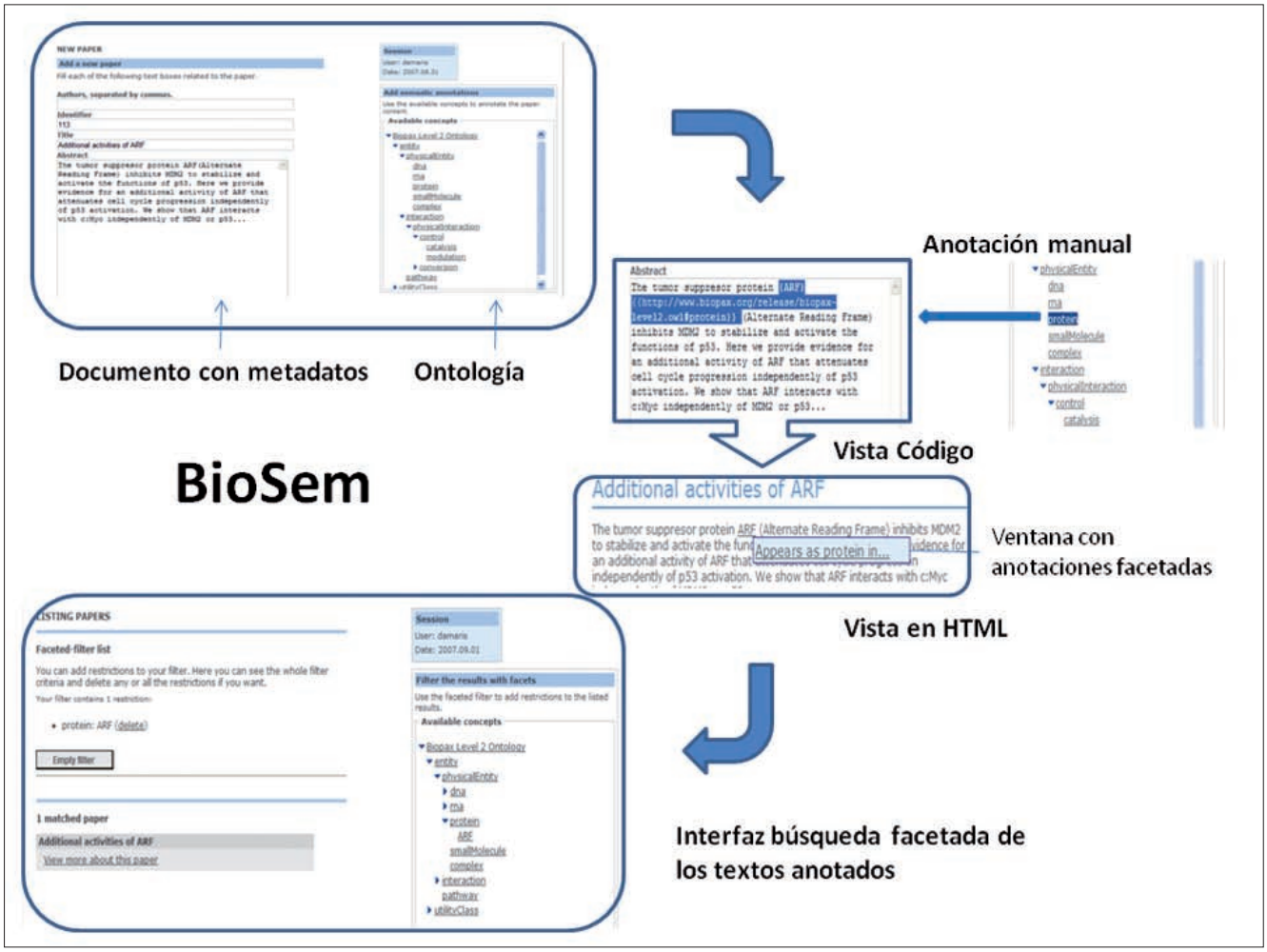

Figura 2. Ejemplo del proceso de anotación mediante la aplicación BioSem (Damaris, Morato y Gómez, 2009)

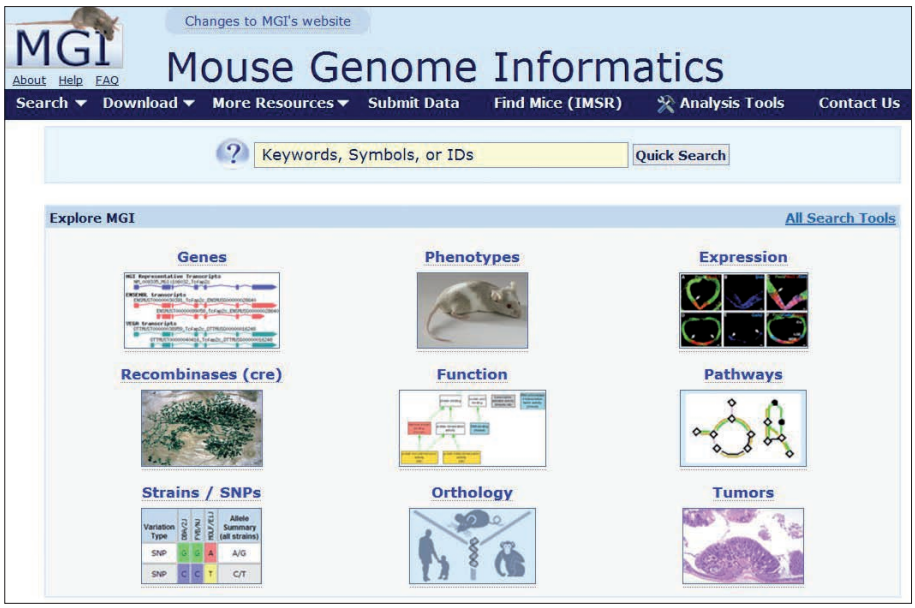

Figura 3. Mouse Genome Informatics, http://www.informatics.jax.org/

con otro conlleva un descenso del 13% en la *F-measure* (**Leser**; **Hakenberg**, 2005). Dado que las herramientas son entrenadas para colecciones particulares, su comportamiento con otras colecciones difiere, y es de esperar que en las aplicaciones reales sea también diferente. Las características terminológicas, entre ellas la existencia de un gran número de palabras compuestas y la necesidad de conocimientos en diversas subáreas, complican la anotación, con tasas de consenso entre anotadores (*inter-annotator agreement*) de 75-90% para genes y proteínas (**Gaizauskas** et al., 2003). Se hace especial empeño en mejorar la consistencia entre anotadores, por ejemplo con la elaboración de un esquema para la anotación semántica en el dominio de la salud pública (**Kawazoe** et al., 2009).

En diversos trabajos se sugiere que la evaluación de los sistemas de recuperación y extracción de información en biomedicina debe orientarse al usuario, creándose métricas y métodos capaces de medir el valor de tales sistemas en aplicaciones reales (**Leser**; **Hakenberg**, 2005; **Cohen**; **Hersch**, 2005). Para ello es necesaria la cooperación entre los expertos en recuperación y extracción de información, y los expertos en el dominio de biomedicina. Ejemplos recientes de este tipo de cooperación incluyen el workshop *BioCreative* 2004 y el *Genomics track* de *TREC*, ambos basados en *gold standards* elaborados por anotadores de bases de datos biológicas en su proceso normal de trabajo.

"Las características terminológicas y la necesidad de conocimientos en diversas subáreas complican las tareas de anotación de corpus para evaluación"

Conclusiones

Son diversas las peculiaridades del dominio biomédico que implican dificultades añadidas para los sistemas de IR. Los problemas de consenso terminológico, tanto en nomenclatura como en organización, y la práctica ausencia de patrones en la terminología utilizada son dos de las más relevantes.

El primer problema influye en la construcción e integración de sistemas de representación de conocimiento y su aplicación en los sistemas de IR. Ante esta situación se impone el uso de internet y la estandarización de los recursos, con formatos y reglas de construcción.

El segundo problema da lugar a un límite en la efectividad de las técnicas basadas en aprendizaje automático. Las iniciativas para normalizar la terminología y las metodologías para lograr consistencia en el etiquetado aportan uniformidad

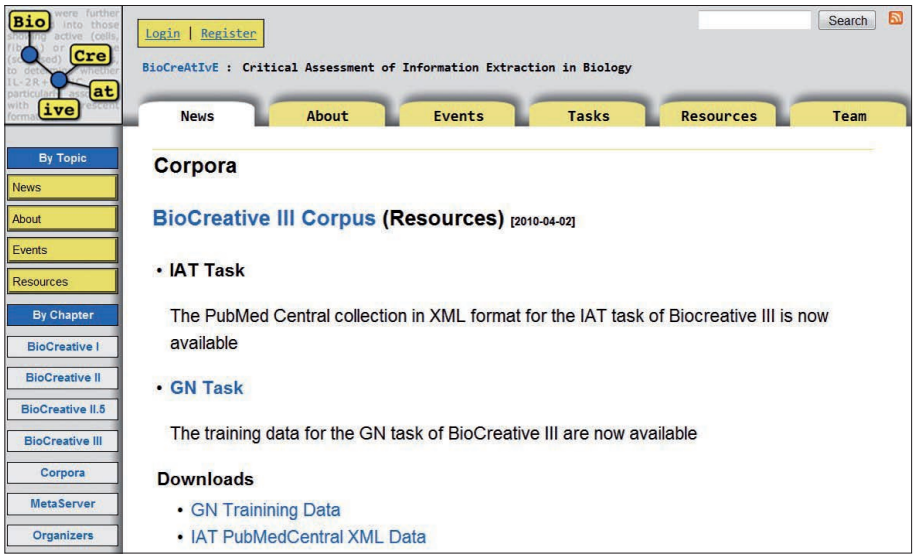

Figura 4. Biocreative, http://www.biocreative.org

Figura 5. Image Clef, http://ir.shef.ac.uk/imageclef

a la información biomédica y contribuyen a mejorar las tasas de efectividad. Es el caso de *The Human Genome Organization* (*HUGO*) con los nombres de genes y proteínas. Estas medidas favorecen además el uso efectivo de las herramientas de anotación semántica como soporte al etiquetado de corpus, lo que supone un medio para lograr una actualización constante de los recursos y proporcionar el soporte que necesitan los sistemas de IR biomédica.

http://www.hugo-international.org

Bibliografía

Ananiadou, Sophia (ed.); **McNaught, John** (ed.). *Text mining for biology and biomedicine*. Artech House, 2006, ISBN 978-1-58053-984-5.

Baeza-Yates, Ricardo. "Tendencias en minería de datos de la Web". *El profesional de la información*, 2009, v. 18, n. 1, pp. 5-10.

Bodenreider, Olivier. "Lexical, terminological and ontological resources for biological text mining". En: Ananiadou, Sophia (ed.); McNaught, John (ed.). *Text mining for biology and biomedicine*. Artech House, 2006, pp. 43-66, ISBN 978-1-58053-984-5.
http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2006/pub2006007.pdf

Clegg, Andrew B.; **Shepherd, Adrian J.** "Evaluating and integrating treebank parsers on a biomedical corpus". En: *Workshop on software (43rd Annual meeting of the ACL)*, 2005.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.8412&rep=rep1&type=pdf

Cohen Aaron; **Hersch, William**. "A survey of current work in biomedical text mining". *Briefings in bioinformatics*, 2005, v. 6, n. 1, pp. 57–71.
http://bib.oxfordjournals.org/cgi/content/short/6/1/57

Collier, Nigel; **Kawazoe, Ai**; **Jin, Lihua**; **Shigematsu, Mika**; **Dien, Dinh**; **Barrero, Roberto A.**; **Takeuchi, Koichi**; **Kawtrakul, Asanee**. "A multilingual ontology for infectious disease surveillance: rationale, design and challenges". *Language resources and evaluation*, 2007, v. 40 n. 3-4, pp. 405-413.
http://naist.cpe.ku.ac.th/downloads/publications/2007_n/Journal_Lecture_Notes/Multi_Onot_Disease.pdf

Cunningham, Hamish. "Information extraction, automatic". En: Brown, Keith (ed.). *Encyclopedia of language and linguistics*, v. 1-14, 2nd Edition, Elsevier Science Publishers, 2005, pp. 665-677. ISBN 0-08-044299-4.
http://gate.ac.uk/sale/ell2/ie/main.pdf

Dingare, Shipra; **Finkel, Jenny**; **Nissim, Malvina**; **Manning, Christopher**; **Grover, Claire**. "A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations". En: *BioLink meeting at ISMB*, 2004.

Gaizauskas, Robert; **Demetriou, George**; **Artymiuk, Pete J.**; **Willett, Peter**. "Protein structures and information extraction from biological texts: the Pasta system". *Bioinformatics*, 2003, v. 19, n. 1, pp. 135–143.
http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/1/135

Hersh, William. *TREC genomics track protocol*. Oregon Health & Science University, 2004.
http://ir.ohsu.edu/genomics/2004protocol.html

Jacquemin, Christian. *Spotting and discovering terms through natural language processing*. Cambridge, MA: MIT Press, 2001, ISBN 0-262-10085-1.

Kawazoe, Ai; **Jin, Lihua**; **Shigematsu, Mika**; **Bekki, Daisuke**; **Barrero, Roberto**; **Taniguchi, Kiyosu**; **Collier, Nigel**. "The development of a schema for semantic annotation: gain brought by a formal ontological method". *Applied ontology*, 2009, v. 4, n. 1, pp. 5-20.

Leser, Ulf; **Hakenberg Jörg**. "What makes a gene name? Named entity recognition in the biomedical literature". *Briefings in bioinformatics*, 2005, v. 6, n. 4, pp. 357-369.

Liu, Hongfang; **Johnson, Stephen**; **Friedman, Carol**. "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS". *Journal of the American Medical Informatics Association*, 2002, v. 9, n. 6, pp. 621–636.

McCray, Alexa T.; **Browne, Allen C.**; **Bodenreider, Olivier**. "The lexical properties of the Gene ontology (GO)". En: *Proceedings of the AMIA symposium*, 2002, pp. 504-508.
http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2002/pub2002030.pdf

Morgan, Alexander; **Hirschman, Lynette**; **Colosimo, Marc**; **Yeh, Alexander**; **Colombe, Jeff**. "Gene name identification and normalization using a model organism database". *Journal of biomedical informatics*, 2004, v. 37, n. 6, pp. 396–410.

Poibeau Thierry; **Kosseim, Leila**. "Proper name extraction from non-journalistic texts". *Language and computers*, 2001, v. 37, pp. 144-157.

Rector, Alan; **Stevens, Robert**; **Rogers, Jeremy**. *Simple bio upper ontolgy*, 2006.
http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/

Rong, Xu; **Morgan, Alex**; **Das, Amar K.**; **Garber, Alan**. "Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon". En: *BioNLP workshop*, 2009, pp. 63-70.
http://aclweb.org/anthology/W/W09/W09-1308.pdf

Rosse, Cornelius; **Kumar, Anand**; **Mejino Jose L. V.**; **Cook, Daniel L.**; **Detwiler, Landon T.**; **Smith, Barry**. "A strategy for improving and integrating biomedical ontologies". En: *Annual symposium of the AMIA*, 2005, pp. 639–643.
http://ontology.buffalo.edu/bio/OBR.pdf

Samwald, Matthias; **Adlassnig, Klaus-Peter**. "The bio-zen plus ontology". *Applied ontology*, 2008, v. 3, n. 4, pp. 213-217.

Schulze-Kremer, Steffen. "Adding semantics to genome databases: towards an ontology for molecular biology". En: *5th Int. conf. on intelligent systems for molecular biology*, 1997, pp. 272-275.

Soldatova, Larisa N.; **King, Ross D.** "Are the current ontologies in biology good ontologies?". *Nature biotechnology*, 2005, v. 23, n. 9, pp. 1095–1098.

Spasic, Irena; **Ananiadou, Sophia**. "A flexible measure of contextual similarity for biomedical terms". En: *Pacific symposium on biocomputing*, 2005, pp. 197-208.
http://helix-web.stanford.edu/psb05/spasic.pdf

Spasic, Irena; **Ananiadou, Sophia**; **McNaught, John**; **Kumar, Anand**. "Text mining and ontologies in biomedicine: making sense of raw text". *Briefings in bioinformatics,* 2005, v. 6, n. 3, pp. 239–251. *http://bib.oxfordjournals.org/cgi/content/short/6/3/239*

Stenzhorn, Holger; **Schulz, Stefan**; **Beißwanger, Elena**; **Hahn, Udo**; **Van Den Hoek, László**; **Van Mulligen, Erik**. "BioTop and ChemTop – Top-Domain ontologies for biology and chemistry". En: *International Semantic Web Conference* (Posters & Demos), 2008, pp. 1-2. *http://www.imbi.uni-freiburg.de/ontology/biotop/publications/iswc08.pdf*

Tsuruoka, Yoshimasa; **Tsujii, Jun'ichi**. "Improving the performance of dictionary-based approaches in protein name recognition". *Journal of biomedical informatics*, 2004, v. 37, n. 6, pp. 461–470.

Weeber, Marc; **Klein, Henny**; **Aronson, Alan R.**; **Mork, James G.**; **De Jong-Van den Berg, Lolkje**; **Vos, Rein**. "Text-based discovery in biomedicine: the architecture of the DAD-system". En: *AMIA symposium*, 2000, pp. 903–907. *http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2000/pub2000061.pdf*

Zhou, GuoDong; **Zhang, Jie**; **Su, Jian**; **Shen, Dan**; **Tan, ChewLim**. "Recognizing names in biomedical texts: a machine learning approach". *Bioinformatics*, 2004, v. 20, n. 7, pp. 1178–1190. *http://bioinformatics.oxfordjournals.org/cgi/content/short/20/7/1178*

Mónica Marrero, Sonia Sánchez-Cuadrado, Julián Urbano, Jorge Morato, Jose-Antonio Moreiro. *Universidad Carlos III de Madrid.*
mmarrero@inf.uc3m.es
ssanchec@ie.inf.uc3m.es
jurbano@inf.uc3m.es
jorge@ie.inf.uc3m.es
jamore@bib.uc3m.es