

TMT: A tool to guide users in finding information on clinical texts

TMT: Una herramienta para guiar a los usuarios en la búsqueda de información sobre textos clínicos

Fernando Aparicio¹, Manuel de Buenaga¹, Margarita Rubio², María Asunción Hernando², Diego Gachet¹, Enrique Puertas¹, Ignacio Giráldez¹

¹ Departamento de Sistemas Informáticos y Automática, Escuela Politécnica

² Departamento de Especialidades Médicas, Facultad de Ciencias Biomédicas

Universidad Europea de Madrid

C/ Tajo, s/n. Villaviciosa de Odón.

{fernando.aparicio,buenaga,margarita.rubio,masuncion.hernando,

Gachet,enrique.puertas,ignacio.giraldez@uem.es }

Resumen: La gran cantidad de información médica disponible a través de internet, tanto en formato estructurado como en formato texto, hace que los distintos tipos de usuario se encuentren con diferentes problemas a la hora de efectuar una búsqueda efectiva. Por un lado, los estudiantes de medicina, el personal sanitario y los investigadores en el área de la biomedicina disponen de una gran variedad de fuentes y herramientas de características dispares, que precisan de un periodo de aprendizaje a veces insalvable. Por otro lado, los pacientes, sus familiares y personas que no pertenecen a la profesión médica, se encuentran con el problema añadido que supone no estar suficientemente familiarizados con la terminología médica. En este artículo presentamos una herramienta que permite extraer conceptos médicos relevantes presentes en un texto clínico, haciendo uso de técnicas para el reconocimiento de entidades nombradas, aplicadas sobre listas de conceptos, y técnicas de anotación a partir de ontologías. Para proponer los conceptos se hace uso de un recurso no formal de conocimiento, como es *Freebase*, y de recursos formales como son *Medlineplus* y *Pubmed*. Nosotros argumentamos que la combinación de estos recursos, con información menos formal y en lenguaje más divulgativo (como es *Freebase*), con información formal y en lenguaje más divulgativo (como es *Medlineplus*) o con información formal y en lenguaje más especializado (como son las publicaciones científicas de *Pubmed*), optimiza el proceso de localización de información médica sobre un caso clínico complejo a usuarios con diferentes perfiles y necesidades, tales como son los pacientes, los médicos o los investigadores. Nuestro objetivo último es la construcción de una plataforma que permita albergar diferentes técnicas para facilitar la práctica de la medicina traslacional.

Palabras clave: Bioinformática, medicina traslacional, ontologías, reconocimiento de entidades nombradas, recuperación de información, aprendizaje basado en casos.

Abstract: The large amount of medical information available through the Internet, in both structure and text formats, makes that different types of users will encounter different problems when they have to carry out an effective search. On the one hand, medical students, health staff and researchers in the field of biomedicine have a variety of sources and tools of different characteristics which require a learning period sometimes insurmountable. On the other hand, patients, family members and people outside of the medical profession, face the added problem of not being sufficiently familiarized with medical terminology. In this paper we present a tool that can extract relevant medical concepts present in a clinical text, using techniques for named entity recognition, applied on lists of concepts, and annotation techniques from ontologies. To propose these concepts, our tool makes use of a non formal knowledge source, such as *Freebase*, and formal resources such as *MedlinePlus* and *PubMed*. We argue that the combination of these resources, with information less formal and more plain language (like *Freebase*), with formal information and more plain language (like *Medlineplus*) or with formal

information and more technical language (such as the Pubmed scientific literature), optimize the process of discover medical information on a complex clinical case to users with different profiles and needs, such as are patients, doctors or researchers. Our ultimate goal is to build a platform to accommodate different techniques facilitating the practice of translational medicine.

Keywords: Bioinformatics, translational medicine, ontologies, NER, information retrieval, case-based learning.

1 Introducción

En un esfuerzo por aumentar la calidad de la medicina, ha emergido recientemente un paradigma denominado *medicina traslacional*, que aplica a las ciencias biomédicas la idea de trasladar eficientemente el conocimiento desde las ciencias básicas hasta el paciente, de manera que se facilite la creación de nuevos tratamientos, asegurando una correcta aplicación de los mismos (Woolf, 2008). Se están desarrollando estudios, desde diferentes puntos de vista, en los que se pretende dilucidar la importancia de la informática en relación a diferentes aspectos vinculados con esta traslación del conocimiento. Por ejemplo, en (Sarkar, 2010) se estudian las diferentes barreras que debe superar la investigación traslacional y el importante papel que tiene la bioinformática para facilitar el intercambio de conceptos entre los especialistas de las diferentes disciplinas involucradas. Otro ejemplo lo podemos encontrar en (Bellazzi et al., 2010), donde se subraya este papel refiriéndose a la bioinformática traslacional y su aportación para la práctica de la medicina personalizada, a través de sistemas de ayuda a la decisión trabajando sobre casos clínicos.

A pesar de la gran cantidad de documentación médica disponible a través de Internet, todavía son muchas las dificultades que el usuario tiene que superar a la hora de buscar este tipo de información, tanto si se trata de un perfil médico como si no. Cuando un paciente, familiar o alguien externo a la profesión médica quiere averiguar más información sobre una posible enfermedad, un conjunto de síntomas o un tratamiento, se encuentra con diversas dificultades entre las que podemos destacar (Luo y Tang, 2008): 1) falta de experiencia y conocimiento del dominio para realizar una búsqueda adecuada, 2) falta de comprensión de la terminología médica encontrada. Por otro lado, cuando un médico o estudiante de medicina trata de encontrar información sobre un caso clínico, por ejemplo, se encuentra con una gran diversidad de

herramientas de búsqueda que dan acceso a diferentes fuentes de información (Trivedi, 2009). Esta variabilidad dificulta el proceso de comprensión, generando una dispersión del proceso y entorpeciendo el proceso de recopilación de información concreta para la resolución del caso de estudio.

Las ontologías biomédicas, tal y como se menciona en (Jonquet, Musen y Shah, 2010), son un elemento cada vez más utilizado a la hora de afrontar tareas de anotación en las investigaciones de biomedicina, permitiendo así una mejor integración de los datos y favoreciendo los descubrimientos traslacionales. Sin embargo, debido a la gran cantidad de ontologías y la heterogeneidad de formatos disponibles, tanto decidir qué ontologías utilizar como ensayar con las seleccionadas se convierten en tareas con un alto coste en tiempo y recursos. Si entre los objetivos incluyésemos el desarrollo de algún tipo de herramienta para procesar estos datos, habría que agregar, al coste anterior, el coste de selección del software y las librerías adecuados para el aprovechamiento de los distintos recursos ontológicos.

En los últimos años, se ha venido produciendo un gran desarrollo de las interfaces de usuario para el acceso a la información de forma amigable y con contenido semántico a través de la Web, técnicas que quedan recogidas en el concepto de Web semántica. Existe una relación entre los objetivos de personalización y recopilación de datos procedentes de diferentes fuentes, inherentes a la investigación traslacional, y los correspondientes en el ámbito de la Web semántica. Han sido realizadas valoraciones de la utilidad de esta última en relación a los objetivos de la primera, como por ejemplo en (Ruttenberg et al., 2009), así como creado grupos de trabajo que estudian campos de aplicación (i.e. HCLSIG¹, Semantic Web for Health Care and Life Sciences Interest Group). La utilidad de los recursos ontológicos en el ámbito de la Web semántica aplicados al dominio biomédico, ya ha sido demostrada en

¹ www.w3.org/2001/sw/hcls/

el desarrollo de otras plataformas, como por ejemplo (Mirhaji et al., 2009).

En este artículo presentamos una herramienta que integra fuentes de conocimiento substanciales, tales como son Freebase² (Bollacker et al., 2008), Medlineplus³ y servicios de anotación con ontologías elaboradas por el National Center for Biomedical Ontology (NCBO⁴). El objetivo es proporcionar información sobre algunos tipos de conceptos médicos presentes en un texto clínico de entrada, ayudando a diferentes perfiles de usuario a la comprensión de su contenido. Para lograrlo, inicialmente se muestra la lista de conceptos detectados en el texto clínico introducido y la fuente con la que se ha detectado, que nos da también información del tipo de concepto, en concreto: enfermedad, síntoma o tratamiento de Freebase y concepto localizado a partir de la ontología Medlineplus Health Topics. Estos conceptos se enlazan con los clientes de búsqueda desarrollados para obtener información a partir de los servicios ofrecidos por las fuentes mencionadas, mostrándose al usuario la descripción del concepto y algunas de las relaciones semánticas, así como la posibilidad de visitar el contenido original navegando al sitio Web de la fuente. Tanto Freebase, a través de las relaciones entre tópicos que vienen definidas por conjuntos de propiedades directamente vinculadas con su dominio (como por ejemplo, los síntomas o los factores de riesgo asociados a una enfermedad, en el dominio de la medicina), como Medlineplus, a través de los sinónimos MeSH o las ontologías utilizadas para la anotación, ofrecen esta posibilidad de tratar las descripciones de los términos junto con un contenido semántico asociado.

El resto del artículo está organizado del siguiente modo. En la sección 2 se diferencia entre dos tipos de información médica relevante, la clínica y la científica, haciéndose mención de las utilizadas para el desarrollo de la herramienta. En la sección 3 se describe la arquitectura del sistema propuesto y los objetivos de los distintos módulos en los que se han agrupado las funcionalidades. En la sección 4 se discute el procedimiento implementado para el reconocimiento de entidades nombradas,

a partir de la recuperación de información desde Freebase. En la sección 5 se presenta la técnica de anotación a partir de ontologías haciendo uso de servicios Web, que eliminan gran parte de la complejidad asociada a este método de procesamiento de texto. En la sección 6 son presentadas las dos posibilidades de acceso que se han desarrollado, ejemplificándose el uso de la herramienta a través del interfaz Web, partiendo de un caso clínico concreto.

2 Fuentes de información en el dominio médico

Se están llevando a cabo iniciativas muy importantes entre cuyos objetivos se incluye el de ofrecer recursos orientados a distintos perfiles de usuario: el público en general, el personal médico o los investigadores. Un ejemplo notable de estos esfuerzos es el llevado a cabo por el National Institutes of Health (NIH⁵), que ofrece una amplia gama de servicios y herramientas bien diferenciadas en función de los tipos de usuario mencionados. Uno de los recursos ofrecido para el público en general es el buscador para problemas de salud Medlineplus, mientras que un recurso ofrecido para el personal médico y los investigadores es el repositorio de literatura biomédica accesible a través de PubMed⁶.

En líneas generales, los distintos tipos de fuentes de información biomédica disponibles se pueden agrupar en información clínica y científica. Durante el desarrollo de este artículo se han hecho uso de las siguientes:

1) Bases de datos de información clínica. En esta categoría se incluyen los siguientes dos tipos. (a) Información sobre conceptos médicos, que puede ser menos formal, como la que contiene Freebase (descrita en la sección 4), o más formal, como la que contiene Medlineplus. (b) Bases de datos de casos clínicos, tales como la base de datos MIMIC-II (Clifford, Scott y Villarreal, 2010), los casos clínicos del departamento de patología de la escuela de medicina de la Universidad de Pittsburgh⁷ o los de la escuela de medicina de Penn State⁸.

2) Bases de datos de publicaciones científicas. Entre las que destaca *PubMed* como una de las fuentes más utilizadas, como dato

² www.freebase.com

³ www.medlineplus.gov

⁴ www.bioontology.org

⁵ www.nih.gov

⁶ www.ncbi.nlm.nih.gov/pubmed

⁷ path.upmc.edu/cases.html

⁸ www.hmc.psu.edu/ume/fcmonline

representativo el número de consultas en 2009 fue de 1300 millones (Shariff et al, 2010).

3 Arquitectura del sistema

Se ha puesto especial énfasis en la construcción de una arquitectura de software que permita albergar y presentar online nuevas fuentes de información con las que enriquecer el soporte a los distintos tipos de usuario, así como la incorporación de técnicas o librerías para el procesamiento del texto y la generación de interfaces de usuario para otro tipo de clientes (como serían los dispositivos móviles u otros aplicativos que necesiten acceder a través de servicios Web). Este diseño de arquitectura para el sistema tiene como objetivo principal facilitar la integración de distintos componentes agrupándolos en módulos. Hasta el momento se han definido los siguientes módulos:

a) Acceso: Donde se agrupan los diferentes mecanismos de acceso al sistema a través del protocolo HTTP, tales como el que se produce al interactuar desde el interfaz Web de usuario o el ofrecido a través de servicios Web. Se encarga de la comunicación con los módulos dedicados al procesamiento del lenguaje natural y a la búsqueda de información asociada a un concepto.

b) Procesamiento del lenguaje natural (PLN): Se nutre de las listas de conceptos obtenidas para procesar el texto utilizando las librerías de GATE⁹ (General Architecture for Text Engineering). También hace uso de los servicios de anotación externos del NCBO.

c) Recuperación de información (RI): Gestiona la descarga de las listas de conceptos, fuera del tiempo de ejecución online, para su uso en el módulo NLP, en tiempo de ejecución.

d) Búsqueda: Proporciona al sistema un interfaz en el que agregar diferentes fuentes para la búsqueda de información, asociada a los conceptos reconocidos en el texto de entrada.

En la Figura 1, se puede encontrar el esquema de la arquitectura y los flujos de información generados al producirse la interacción del usuario a través del interfaz Web, que también se corresponden con los flujos generados a través de la interacción con el servicio Web del módulo de acceso.

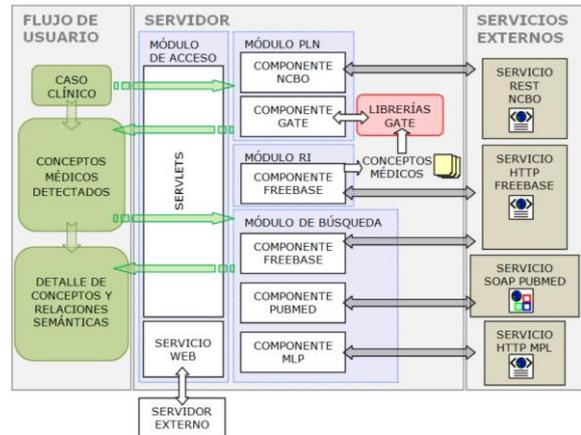


Figura 1: Arquitectura del sistema y flujo de información

4 Reconocimiento de Entidades Nombradas con Freebase

El sistema propuesto pre-procesa un conjunto de conceptos médicos recopilados en listas, que posteriormente son utilizados por el componente *Gazetteer* de sistema *ANNIE* incluido en la distribución de GATE. Para la obtención de estas listas de conceptos médicos se ha utilizado Freebase, que es una base de datos colaborativa de conocimiento publicada en 2007 por Metaweb y adquirida por Google recientemente. Freebase estructura el contenido basándose en los denominados *tópicos*, *tipos*, *dominios* y *propiedades*. Los tópicos agrupan la información descriptiva asociada a una entidad nombrada, se corresponden típicamente con los artículos de Wikipedia, y mantienen una relación semántica con los tipos y las propiedades: El tópico "es_un" tipo o "tiene_una" propiedad. Los tipos son agrupados en dominios (i.e. medicine, astronomy, etc.) a los que se les asigna un identificador relacionado con el enlace al contenido (/medicine, /astronomy, etc.) y contienen un conjunto de propiedades, que nos darán relaciones para todos los tópicos de ese tipo. A los tipos también se les asigna un identificador, que se construye a partir del identificador del dominio. Por ejemplo, si se considera el tópico *cáncer*. Este tópico es un tipo de *enfermedad* (perteneciendo al dominio medicina con un identificador /medicine/disease) y también es un tipo de *constelación* (formando parte del dominio astronomía con un identificador

⁹ gate.ac.uk

/astronomy/constellation). La información de tópicos relacionados se puede obtener a partir de las propiedades de cada tipo, por tanto si se selecciona el tópico cáncer del tipo enfermedad se dispondría de propiedades como: síntomas, tratamientos, factores de riesgo, etc. Por otro lado, si se considera que es del tipo constelación, se tendrían propiedades como: estrellas, constelaciones limítrofes, lluvias de meteoritos, etc. Además, las relaciones de una propiedad, como por ejemplo los síntomas asociados al tipo enfermedad en el dominio médico, pueden estar definidas en otro tipo bajo el mismo dominio disponiendo de su propio identificador, es decir, en el ejemplo tendríamos un tipo del dominio médico denominado síntoma (su identificador sería /medicine/symptom), que a su vez posee sus propias propiedades: efecto secundario de, síntoma de, etc.

A pesar de no ser una ontología formal de conocimiento, la calidad de la información es alta precisamente por la introducción manual, tal y como se menciona en (Han y Zhao, 2009), donde se utiliza Freebase para desambiguar nombres por sus profesiones. También ha sido utilizado en otros ámbitos, como son los trabajos relacionados con utilidades en el marco de la Web 2.0 y 3.0 (Auer S. et al, 2007; Bizer C. et al, 2009), clasificaciones de consultas (Brenes, Avello y González, 2009), como extensiones al geo-etiquetado (Weichselbraun, 2009) o como ejemplo de la navegación por facetas (Hearst, 2008).

Uno de los métodos de acceso a la información que provee la base de datos es lo que se denomina Metaweb Query Language, en (Meyer et al., 2010), disponiéndose para ello de diferentes APIs de acceso dependiendo del tipo de lenguaje de programación utilizado. Como herramienta de pruebas, Freebase ofrece un editor público¹⁰. A lo largo de nuestro desarrollo se ha utilizado el acceso a los datos con AJAX y con Java. La arquitectura final ha mostrado un comportamiento más estable usando Java, tanto para el pre-procesado como para la búsqueda de los tópicos y de las relaciones semánticas.

5 Anotación con ontologías

El NCBO está construyendo herramientas y servicios de código libre para dar soporte al

trabajo de la comunidad biomédica con ontologías. Uno de estos recursos es el denominado BioPortal (Musen et al., 2008, Noy et al., 2009), un repositorio de ontologías biomédicas en diferentes formatos, que ofrece la posibilidad de acceso a través de servicios o directamente a través de un interfaz Web¹¹. Estos servicios permiten focalizar el esfuerzo de la investigación biomédica sobre los datos haciendo uso de multitud de ontologías, con estructuras complejas, en tareas relacionadas con el procesamiento del lenguaje natural (72 ontologías en Marzo de 2008, 134 en Mayo de 2009 y 202 en el momento de redacción de este artículo, Diciembre de 2010). Además del acceso unificado a una gran variedad de ontologías, dispone de un conjunto de recursos indexados de forma automática a partir de dichas ontologías (Open Biomedical Resources, OBR), que también poseen las mismas facilidades de acceso por medio de interfaz o servicios Web. En Diciembre de 2010 el número de recursos indexados es de 22. Otro servicio publicado recientemente es el *Recommender* (Jonquet, Musen y Shah, 2010), que propone una clasificación de ontologías con las que anotar el texto, basada en criterios de cobertura, conectividad y número de apariciones del concepto en cada ontología.

En el desarrollo de este artículo se han estudiado los resultados de algunas de estas ontologías y su tratamiento a través del servicio NCBO Open Biomedical Annotator (OBA), descrito en (Jonquet, Shah y Musen, 2009). Este servicio ya ha sido utilizado anteriormente en otros sistemas como son los mencionados en (Dowell et al., 2009), donde se realiza la evaluación de varias de las ontologías para su integración con un desarrollo de los autores, o en (Twigger, Geiger y Smith, 2009), donde se utiliza el OBA para la anotación de registros del repositorio de datos Gene Expression Omnibus. En nuestro sistema se ha utilizado la ontología *Medlineplus Health Topics*, que ha permitido la obtención de información asociada a los conceptos a partir de los servicios Web proporcionados por Medineplus¹².

¹⁰ www.freebase.com/app/queryeditor

¹¹ bioportal.bioontology.org

¹²

www.nlm.nih.gov/medlineplus/webservices.html

6 Métodos de acceso y elementos de la interfaz

Se han desarrollado dos mecanismos de acceso: (1) A través de un interfaz Web¹³, que permite al usuario operar con el texto clínico desde su navegador. (2) Por medio de servicios Web, posibilitando la obtención de los resultados desde otros sistemas.

Para realizar las evaluaciones previas a la evaluación sistemática de la herramienta, se han hecho uso de historiales médicos procedentes de MIMIC-II y casos clínicos facilitados por el Penn State College of Medicine (casos 20¹⁴ y 21¹⁵) y por el departamento de patología de la Universidad de Pittsburgh (casos elegidos al azar: 223, 410, 474, 564, 565 y 616). Para ejemplificar el uso de la herramienta a continuación se utilizará uno de estos últimos casos clínicos, que son ampliamente utilizados en medicina y otros dominios (i.e. Watson, Smith y Watter, 2005, Ruiz-Rico, Vicedo y Rubio-Sánchez, 2008), formando parte de las referencias manejadas por docentes de la Facultad de Ciencias Biomédicas de nuestra Universidad. Estos casos poseen una estructura en la que se divide el texto, en líneas generales, en una historia previa, los resultados de las pruebas y el diagnóstico final. En concreto, se ha seleccionado el caso 223¹⁶, que es el caso de una mujer de 37 años con un diagnóstico de miocardiopatía arritmogénica del ventrículo derecho, con muchos indicios debido a la gran cantidad de patologías relacionadas en su familia. El procesado del texto completo de este caso da como resultado un amplio conjunto de términos de interés, que proporcionan una idea previa del problema o del diagnóstico, dado que muchos de ellos están relacionados con patologías cardiacas. A continuación se muestra el comportamiento del sistema para el primer párrafo de la historia clínica de esta persona, que se puede encontrar en la zona izquierda del interfaz mostrado en la Figura 2.

Tras introducir el texto de este primer párrafo en la herramienta y procesarlo, se muestran las entidades médicas reconocidas y las fuentes, en una tabla que permite ordenar y filtrar por concepto y por fuente. El resultado se puede ver en la zona derecha de la Figura 2. El

mismo procedimiento puede ser aplicado al texto completo del caso.

The screenshot shows a web interface with a text area on the left and a table on the right. The text area contains a clinical case description. The table on the right lists concepts extracted from the text and their corresponding sources.

CONCEPT	SOURCE
ATRIAL FIBRILLATION	Freebase Disease
MITRAL REGURGITATION	Freebase Disease
VENTRICULAR TACHYCARDIA	Freebase Disease
TACHYCARDIA	Freebase Disease
ATRIAL FIBRILLATION	Freebase Symptom
VENTRICULAR TACHYCARDIA	Freebase Symptom
TACHYCARDIA	Freebase Symptom
CARDIAC CATHETERIZATION	Freebase Treatment
DIGOXIN	Freebase Treatment
DIURETIC	Freebase Treatment
AMIODARONE	Freebase Treatment
SYMPTOMS	MedlinePlus
TACHYCARDIA	MedlinePlus
ATRIAL FIBRILLATION	MedlinePlus

Figura 2: Resultados de la herramienta tras procesar el texto

Los conceptos resultantes están enlazados con una página que nos muestra información obtenida a partir de los servicios ofrecidos por la fuente. Por ejemplo, el concepto "atrial fibrillation" ha sido anotado con la fuente Freebase, como enfermedad y como síntoma, y también con la fuente Medlineplus (que nos indica que es una anotación realizada a través la ontología del NCBO denominada Medlineplus Health Topics). Esto nos permite obtener una descripción y relaciones semánticas del concepto tratándolo como si se considerase que es un tipo de enfermedad o un tipo de síntoma en Freebase, así como una descripción de los tópicos resultantes en Medlineplus al hacer una búsqueda por este término, que incluye un conjunto de conceptos MeSH para cada tópico. Profundizando en la información de enfermedades de Freebase se puede obtener, además de una descripción, un conjunto de síntomas, factores de riesgo y tratamientos asociados a la enfermedad. Cada uno de estos términos relacionados semánticamente se enlaza al tópico de la fuente. En caso de profundizar en el contenido del tipo síntoma, se puede obtener información semántica a través de las propiedades "efecto secundario de" y "síntoma de". Si el tipo seleccionado es un

¹³ www.esp.uem.es:8080/MedicalFace/mt.html

¹⁴ www.hmc.psu.edu/ume/fcmonline/case20

¹⁵ www.hmc.psu.edu/ume/fcmonline/case21

¹⁶ path.upmc.edu/cases/case223.html

tratamiento, la información relacionada se obtiene a través de las contraindicaciones, efecto secundario y ensayos. Si se selecciona la fuente Medlineplus, además de la descripción se obtienen otros resultados relacionados con la búsqueda, tal y como son Arrhythmia, Blood Thinners y otros, y los sinónimos de MeSH para cada uno de ellos (los términos MeSH para Arrhythmia son Arrhythmias y Cardiac, mientras que para Blood Thinners son Heparin, Platelet Aggregation Inhibitors, Warfarin y Anticoagulants). Ya por último, existe la posibilidad de capturar un listado de publicaciones científicas en PubMed, a partir de la información resultante en las búsquedas sobre Freebase y sobre Medlineplus. En el primer caso, se enlaza el concepto para la búsqueda de publicaciones relacionadas, mientras que en el segundo caso se enlaza el sinónimo MeSH.

7 Conclusiones y trabajo futuro

En este artículo hemos presentado una herramienta, desarrollada en el marco de la medicina traslacional, que agrupa información de fuentes de conocimiento relevantes, como son el sistema Freebase, la ontología Medlineplus Health Topics del NCBO, el buscador de tópicos Medlineplus y el repositorio de literatura científica PubMed, para ofrecer al usuario un acceso integrado a la información.

También hemos mostrado varias formas de enlazar información semántica, a través del sistema Freebase y de Medlineplus, así como para relacionar esta información con el repositorio de literatura científica PubMed. Se ha descrito la arquitectura del sistema y un método de desarrollo sistemático para la integración de nuevas funcionalidades. Finalmente, tras obtener resultados positivos en las evaluaciones previas, hemos ejemplificado su utilidad utilizando un caso clínico.

En trabajos futuros trataremos de incorporar una evaluación sistemática, nuevas estrategias de lingüística computacional, integración de otros recursos y nuevos interfaces de usuario. En relación con la evaluación, estamos preparando una prueba para medir de forma objetiva y subjetiva la utilidad de uso con estudiantes de medicina de segundo curso, basada en el soporte a la comprensión de casos clínicos. Para la aplicación de nuevas funcionalidades lingüísticas, se está

desarrollando el acceso a un servicio que detecta la negación de los conceptos utilizando la frase a la que pertenecen, considerándose este etiquetado de mucha utilidad dado que los conceptos médicos aparecen negados con bastante frecuencia. En relación a los recursos, estamos valorando la integración de nuevos contenidos procedentes de las fuentes ya utilizadas (nuevas listas de Freebase y otras ontologías o recursos del NCBO) y también de nuevas relaciones entre los resultados. Respecto a las nuevas interfaces, estamos estudiando el acceso a la herramienta desde dispositivos móviles y actualmente estamos desarrollando mecanismos para el acceso desde una aplicación de escritorio.

Referencias bibliográficas

- Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak y Z. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. En *The Semantic Web*, Volume 4825 of *Lecture Notes in Computer Science*, Chapter 52, 722-735-735. Berlin (Heidelberg).
- Bellazzi R., C. Larizza, M. Gabetta, G. Milani, A. Nuzzo, V. Favalli y E. Arbustini. 2010. Translational Bioinformatics: Challenges and Opportunities for Case-Based Reasoning and Decision Support. *LNCS: Case-Based Reasoning*, Vol. 6176, pp. 1-11.
- Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak y S. Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web* 7(3):154-165.
- Bollacker, K., C. Evans, P. Paritosh, T. Sturge y J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. En *Proceedings of the international conference on Management of data, ACM SIGMOD*, pp. 1247-1250. Vancouver (Canada).
- Brenes, D. J., D. G. Avello y K. P. González. 2009. Survey and evaluation of query intent detection methods. En *Proceedings of the workshop on Web Search Click Data*, pp. 1-7. Barcelona (Spain).
- Clifford, G., D. J. Scott y M. Villarroel. 2010. User Guide and Documentation for the MIMIC II Database, Rev: 259. Cambridge, MA, USA.

- Dowell, K. G., M. S. McAndrews-Hill, D. P. Hill, H. J. Drabkin y J. A. Blake. 2009. Integrating text mining into the MGI biocuration workflow. *Database* 2009(0):bap019+.
- Han, X. y J. Zhao. 2009. CASIANED: Web Personal Name Disambiguation Based on Professional Categorization. En *Proceedings of 2nd Web People Search Evaluation Workshop, WePS2*, Madrid (Spain).
- Hearst M. A. 2008. UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. En *Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval*, pp. 13-17, Redmond
- Jonquet, C., N. Shah y M. Musen. 2009. The Open Biomedical Annotator. En *AMIA Summit on Translational Bioinformatics*, pp. 56-60, San Francisco (CA).
- Jonquet, C., M. A. Musen y N. H. Shah. 2010. Building a biomedical ontology recommender web service. *Biomedical semantics*, 1(1):S1.
- Luo, G. y C. Tang. 2008. On iterative intelligent medical search. En *Proceedings of the 31st annual international conference on Research and development in information retrieval, ACM SIGIR*, pp. 3-10, Singapore.
- Meyer S., J. Degener, J. Giannandrea y B. Michener. 2010. Optimizing schema-last tuple-store queries in graphd. En *Proceedings of the International Conference on Management of Data, ACM SIGMOD*, pp. 1047-1056, Indianapolis (Indiana).
- Mirhaji, P., M. Zhu, M. Vagnoni, E. Bernstam, J. Zhang y J. Smith. 2009. Ontology driven integration platform for clinical and translational research. *BMC Bioinformatics* 10(Suppl 2):S2+.
- Musen, M., N. Shah, N. Noy, B. Dai, M. Dorf, N. Griffith, J. D. Buntrock, C. Jonquet, M. J. Montegut y D. L. Rubin. 2008. BioPortal: Ontologies and Data Resources with the Click of a Mouse. En *Annual Symposium proceedings, AMIA*, pp. 1223-1224.
- Noy, N. F., N. H. Shah, P. L. Whetzl, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M. A. Storey, C. G. Chute y M. A. Musen. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research* 37 (Web Server issue): W170-173.
- Ruiz-Rico, F., J. Vicedo y M. Rubio-Sánchez. 2008. Multilingual assistant for medical diagnosing and drug prescription based on category ranking. En *22nd International Conference on Computational Linguistics: Demonstration Papers, ACL COLING*, pp. 169-172, Morristown (NJ).
- Ruttenberg, A., J. A. Rees, M. Samwald y M. S. Marshall. 2009. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in bioinformatics* 10(2):193-204.
- Sarkar, I. 2010. Biomedical informatics and translational medicine. *Translational medicine* 8(1):22+.
- Shariff, S., M. Cuerden, R. B. Haynes, K. A. McKibbin, N. Wilczynski, A. Iansavichus, M. Speechley, A. Thind y A. Garg. 2010. Evaluating the impact of MEDLINE filters on evidence retrieval: study protocol. *Implementation Science* 5(1):58+.
- Trivedi, M. 2009. A study of search engines for health sciences. *Library and Information Science*, 1(5): 069-073
- Twigger, S., J. Geiger y J. Smith. 2009. Using the NCBO Web Services for Concept Recognition and Ontology Annotation of Expression Datasets. En *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS*, Vol-559.
- Watson, M., A. Smith y S. Watter. 2005. Leximancer concept mapping of patient case studies. En *LNCS: Part III. Knowledge-based intelligent information and engineering systems 9th international conference*, pp. 1232-1238, Berlin (Heidelberg).
- Weichselbraun, A. 2009. A Utility Centered Approach for Evaluating and Optimizing Geo-Tagging. En *First International Conference on Knowledge Discovery and Information Retrieval, KDIR*, pp. 134-139, Madeira (Portugal).
- Woolf, S. H. 2008. The Meaning of Translational Research and Why It Matters. *JAMA* 299(2):211-213.