Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 29-38 recibido 05-05-12 revisado 28-05-12 aceptado 04-06-12

Sistema de Acceso a la Información basado en conceptos utilizando Freebase en Español-Inglés sobre el dominio Médico y Turístico

Information Access System based on concepts using Freebase in Spanish-English over the domain Medical and Tourist

Rafael Muñoz
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

Fernando Aparicio
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

Manuel de Buenaga
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

{rafael.munoz, fernando.aparicio, buenaga@uem.es}

Resumen: En este artículo presentamos una herramienta de acceso a la información, basado en los conceptos, enfocada tanto a textos médicos como turísticos. Usando técnicas para el marcado de entidades reconocidas, el sistema permite extraer conceptos relevantes para aportar más información sobre ellos utilizando bases de conocimiento colaborativas y ontologías.

Componentes especialmente interesantes para el desarrollo del sistema son Freebase, una gran base de conocimiento colaborativa, además de recursos formales como MedlinePlus y PubMed. La arquitectura del sistema ha sido construida pensando en términos de escalabilidad, para constituir una gran plataforma de integración de información, con los siguientes objetivos: permitir la integración de diferentes técnicas de procesamiento de lenguaje natural y ampliar las fuentes desde las que se extrae información, así como facilitar la integración de nuevas interfaces de usuario.

Palabras clave: Extracción de Información, Integración de Información, Bases de datos Colaborativas, Procesado de Textos, Arquitectura Escalar, Freebase

Abstract: In this paper we present a tool for access to information, based on semantic, focused both medical texts and tourists. Using marking techniques for recognized entities, the system can extract relevant concepts to provide more information about them, using collaborative databases and ontologies.

Particularly relevant components to its the development are Freebase, a large collaborative base of knowledge and formal resources such as MedlinePlus and PubMed. The platform architecture has been built thinking in terms of scalability, in order to constitute a great platform for information integration, with the following objectives: to allow the integration of different natural language processing techniques, to expand the sources from which information extraction can be performed and to ease integration of new user interfaces.

Keywords: Information Extraction, Information Integration, Collaborative Databases, Text Processing, Scalable Architecture, Freebase.

1 Introducción

Los sistemas de acceso a la información se nutren en la actualidad de un número creciente de diversas fuentes externas como son las ontologías, almacenamientos heterogéneos o incluso redes de conocimiento colaborativas (Gutiérrez V. Y. et al, 2010). En los últimos años se ha incrementado notablemente la cantidad de sitios donde poder acceder a información (Chang C. H. et al, 2006). Esto hace que la cantidad de información sea cada vez mayor y más heterogénea, que se encuentre estructurada o desestructurada, haciendo

ISSN 1135-5948

cada vez más necesarios los sistemas de búsqueda y recuperación de información efectivos (Allan J. et al, 2005). Estos evolucionan rápidamente y ya no solo tienen como objetivo que muestren una enorme cantidad de información, sino que la información sea lo más útil posible (Egozi O. et al, 2011). Es por ello que lo que se impone son los sistemas de integración de información (Tuchinda R. et al, 2011), donde se aúnan diversas técnicas de anotación semántica, análisis de la información, recomendación y personalización. Todo esto para facilitar el acceso más que a una gran cantidad de información no relacionada, a información precisa y que le permita ampliarla profundizando en la cadena de búsqueda.

En este artículo presentamos un sistema de acceso a la información basado en la semántica, que sienta las bases del desarrollo de un sistema mayor de integración de información. Los dominios de aplicación escogidos son el biomédico y el turístico. El primero de ellos con el objetivo de conseguir un sistema destinado a introducirse en el campo de la Medicina Personalizada (Hamburg M et al, 2010) y el segundo, seleccionado por la heterogeneidad del vocabulario y la diversidad de formatos de la información almacenada, generalmente multimedia (Lew M. S. et al, 2006) (texto, fotos, video…).

El sistema propuesto en este artículo permite el acceso a la información a través de la identificación de conceptos relevantes y la integración del conocimiento sobre dichos conceptos, almacenados en la base de conocimiento colaborativa Freebase, que son procesados con técnicas lingüísticas computacionales a través del sistema GATE (Cunningham H. 2002). El resultado final es un sistema que ayuda a optimizar el tiempo dedicado a comprender el texto y ampliar fácilmente la información relevante contenida en él.

El resto del artículo está organizado como sigue. La sección 2 trata sobre los asuntos relacionados con la extracción de información de la base de datos Freebase, las ontologías y la indexación de conceptos. La sección 3 está dedicada a la arquitectura del sistema. La Sección 4 describe el procesamiento de información, se ilustra el interfaz Web a través de dos ejemplos de los dominios seleccionados. Finalmente se exponen las conclusiones, así como futuros trabajos.

2 Uso de Freebase en el acceso semántico

Freebase (Bollacker K. et al, 2008) es una gran base de datos de conocimiento colaborativa, publicada en 2007 por Metaweb y recientemente adquirida por Google.

En nuestra propuesta, Freebase es usada para recuperar listas de conceptos médicos o turísticos para el reconocimiento de las entidades mencionadas (Nadeau D. and Sekine S., 2007) en textos y conectarlos con contenidos semánticamente relacionados. Podríamos categorizar a Freebase como una ontología, ya que los términos se encuentran perfectamente clasificados.

La importancia de las ontologías referentes a dominios particulares es ampliamente reconocida. El objetivo que tenemos al utilizar una ontología es reducir al máximo la confusión entre conceptos. En el ámbito médico, por ejemplo, entre profesionales de la medicina, o en el dominio del turismo, entre operadores turísticos (Navigil R. and Velardi P. 2004).

En nuestro sistema combinamos Freebase con ontologías propias de los dominios. Presentamos una propuesta de método de acceso en dos idiomas, inglés y español gracias a la característica bilingüe de Freebase. Esto nos permite introducir los textos en cualquiera de los idiomas, devolviéndonos los resultados en el idioma origen de la fuente.

2.1 Escenarios de aplicación Cross-Lingüe

La recuperación de información cross-lingüe (CLIR) está relacionada con la recuperación de datos en idiomas diferentes al utilizado por el usuario, facilitando el acceso a recursos por otros criterios como la similitud o la calidad. Este tipo de tareas de recuperación es uno de los objetivos destacados en la iniciativa CLEF[1] y en otras de más reciente creación, como las mencionadas en (Mayfield et al., 2011).

[1] http://www.clef-initiative.eu/

Uno de los recursos más utilizados para la recuperación de información cross-lingüe son los contenidos de la Wikipedia, u otros más recientes como el Wiktionary (Müller y Gurevych, 2009). Las ontologías permiten el almacenamiento de datos en diferentes idiomas, conectando los conceptos a través de meta-información. Esto las convierte en una herramienta muy útil en la construcción de sistemas cross-lingüe (Carrero et al., 2007; Knoth et al., 2010). Freebase es un sistema que relaciona ambos ámbitos, es decir, es una ontología construida a partir de la extracción multilingüe de información desde la Wikipedia y otras fuentes (De Melo y Weikum, 2010).

El sistema desarrollado parte de un caso clínico en inglés para, a través del procesamiento del texto, extraer un conjunto de conceptos (en inglés) sobre los que es posible ampliar información (también en inglés). Sin embargo, este es sólo uno de los escenarios posibles. Cada uno de estos elementos puede ser llevado a cabo en inglés o en otros idiomas, como por ejemplo el español.

La generación de listas de conceptos procedentes de Freebase o de Medlineplus en ambos idiomas (en el formato apropiado para su procesado con el elemento Gazetteer de GATE), proveen de información detallada de los conceptos detectados ofrecida en inglés, mientras que los textos de entrada y los conceptos pueden estar en ambos idiomas. Además, dado que Medlineplus también proporciona los contenidos en español a sus usuarios, y que la información detallada de Freebase está fundamentalmente basada en la Wikipedia, es posible acceder a información de detalle mostrada en español.

Estos escenarios bilingües pueden contribuir al enriquecimiento de la metodología de aprendizaje publicada en (Aparicio et al., 2011a), tal y como ya se ha hecho en otras experiencias educativas. Un ejemplo próximo se puede encontrar en (Clark et al., 2012), donde se utiliza este tipo de sistema bilingüe para la realización de una experiencia con alumnos de un centro de educación secundaria en Estados Unidos, obteniéndose resultados muy interesantes en cuanto a la utilidad de la experiencia (ofrecer los recursos en ambos idiomas sirve tanto para facilitar la comprensión de la materia como para la mejora del idioma no nativo).

2.2 Ontologías turísticas y biomédicas

Turísticas: La proliferación de ontologías turísticas ha desembocado en el desarrollo y profundización de las mismas, llevándonos a realizar ontologías de ámbito regional.

Seguidamente se enumeran las ontologías turísticas más representativas a día de hoy:

Harmonise, IMHO (Interoperable Minimum Harmonization Ontology), Hi-Touch, Tourist Ontology por AIFB, Mondeca, Qall-Me. Existen un par de Ontologías Españolas: Cruzar, y ANOTA.

Estas ontologías contienen conceptos del dominio turístico y actividades de ocio, abarcando entidades turísticas, culturales, paquetes turísticos e incluso contenido multimedia.

En el ámbito de la biomedicina existen muchos ejemplos de ontologías desarrolladas a lo largo de los últimos años. Algunos ejemplos destacados, de este amplio conjunto de recursos biomédicos, son: GO (Gene Ontology), UMLS (Unified Medical Language System), SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) o FMA (Foundational Model of Anatomy).

2.3 Freebase

El uso de Freebase en este sistema se debe a que la información almacenada se encuentra estructurada. Esta se agrupa en "dominios" (como medicina, viajes, deportes, etc). Dentro de los dominios existen los "tópicos", que almacenan toda la información de ese concepto y tiene una relación semántica con tipos y propiedades de la forma "es un" tipo o "tiene una" propiedad. Los dominios, tipos y propiedades tienen un identificador único, consistente en la concatenación del tipo o el tópico, como se muestra en los dos siguientes ejemplos. (A) El tópico "autism" pertenece al dominio *medicine* y tipo *disease* de ahí su identificador es /medicine/disease. (B) El tópico "Taj Mahal" pertenece al tipo *tourist attraction* que se encuentra englobado en el dominio *travel*, por lo tanto su identificador es /travel/tourist attraction. Freebase ha sido usado en otras áreas de investigación, como piezas de trabajo relacionado a software en el contexto de la Web 2.0 y 3.0, herramientas para la desambigüación de

nombres (Han X. and Zhao J., 2009), clasificación de consultas (Brenes D. J. et al, 2009).

El sistema consigue listas de *diseases*, *symptoms* y *treatments* del dominio *medicine* y *tourist attraction*, *accommodation*, *travel destination* del dominio *travel* extraído de Freebase para el posterior reconocimiento usando la herramienta GATE (Aparicio F. 2011b).

2.4 Indexación Conceptual

Dependiendo del tipo de documentos a analizar, la indexación conceptual requiere solventar problemas como: desambiguación semántica, relaciones semánticas, traducción automática, resolución de polisemia, etc. (Verdejo F. et al 1999)

Nuestro sistema realiza un proceso de indexación conceptual en el que se procesan los textos y se identifican los tópicos (conceptos) de Freebase que aparecen en ellos, marcándolos para que a través de hipervínculos se pueda profundizar en el conocimiento (Voss A. et al, 1999).

Para el objetivo propuesto para este sistema, existen algunos de aspectos que obtienen mayor relevancia. Así, contar con un número de términos representativo (14.000 médicos y más de 4.000 turísticos) en los que profundizar en el conocimiento, y que sea una aplicación plenamente funcional utilizada por usuarios reales permitiéndoles el reconocimiento de entidades nombradas es el objetivo propuesto y conseguido.

Para este proceso de textos orientado a la indexación conceptual, utilizamos como elemento principal el sistema GATE (Generic Architecture for Text Engineering) (Cunningham H. et al, 2002).

Entre las diferentes opciones de programación para integrar en este software, hemos seleccionado una que admite el desarrollo y prueba con la GUI y reutiliza la lógica del módulo NLP. GATE es distribuido incluyendo un sistema de extracción de información llamado ANNIE (A Nearly-New Information Extraction System), incorporando un amplio rango de recursos que realizan tareas del análisis del lenguaje a diferentes niveles.

Gazetteer es uno de sus componentes, al que se le ha dado una importancia especial en nuestro diseño del sistema. Este componente, basado en listas predefinidas, permite el reconocimiento de entidades previamente mencionadas. Estas listas, a su vez, permiten la inclusión de detalles de cada entidad, que en nuestro caso son principalmente usados para almacenar los identificadores Freebase.

2.5 Arquitectura

Con respecto a la arquitectura del sistema, se ha puesto un particular énfasis en la creación de una arquitectura de software habilitando el almacenamiento y la presentación online de nuevas fuentes de información. Esto enriquecerá la interface del usuario, así como la incorporación de nuevas técnicas en el procesamiento de palabras y la generación de interfaces de usuario para diferentes tipos de clientes tales como dispositivos móviles u otros que requieran acceso a través de servicios web. El principal objetivo de este diseño del sistema es hacer la integración de diferentes componentes fácilmente agrupándolos en diferentes módulos. La modularización de componentes tiene los siguientes objetivos específicos:

- Acceso: recopila los diferentes mecanismos para acceder al sistema a través del protocolo HTTP, tal como el producido por interactuar desde la interfaz web del usuario o el ofrecido a través de servicios web. Es el responsable de la comunicación entre aquellos módulos dedicados a procesamiento del lenguaje natural y los de búsqueda de información asociada a un concepto particular.
- El módulo de procesamiento de lenguaje natural habilita al sistema a usar otras herramientas o librerías, manteniendo la misma estructura y haciendo posible la interacción entre diferentes componentes. Igualmente, hace uso de la lista de conceptos obtenidos a procesar texto usando librerías GATE.
- El módulo de recuperación de información extrae la información en tiempo de ejecución del sistema, optimizando el tiempo de respuesta online.
- Búsqueda: provee al sistema con un interfaz para que diferentes fuentes de búsqueda de información, asociadas a conceptos que aparecen en textos de entrada, puedan ser añadidos.

La Figura 1 muestra la arquitectura y el flujo de información generado cuando el usuario interactúa a través del interfaz Web.

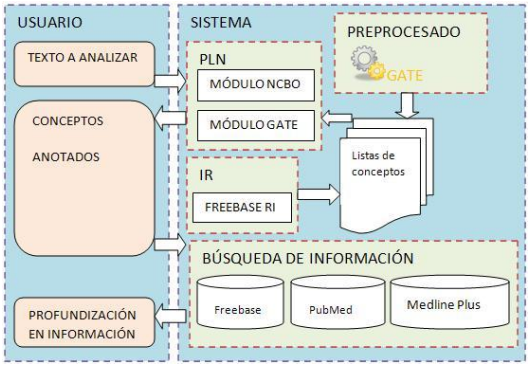

Figura 1: Arquitectura del sistema

3 Casos de Uso

Hemos escogido dos dominios diferentes, cada uno con ciertas características propias, en los que aplicar nuestro sistema, ambos ampliamente utilizados, el dominio biomédico y el dominio turístico.

Los textos biomédicos tienen ciertas características que los diferencian de los textos de otras disciplinas. Las peculiaridades de terminología y estilo de escritura usado por los profesionales médicos hacen la detección de conceptos y el análisis de los datos una tarea compleja y ambiciosa.

En el caso del turismo, el léxico utiliza una amplia terminología de varios campos (geografía, economía, arte, historia, etc.). Es, por lo tanto, un lenguaje poco especializado, que usa un vocabulario amplio y habitual.

Se han creado dos mecanismos de acceso a nuestro sistema: (1) A través de un interfaz web, permitiendo al usuario operar los textos desde su navegador. (2) A través de servicios web, permitiendo la recuperación de resultados de otros sistemas.

3.1 Aplicación al Dominio Biomédico

La búsqueda de información sobre salud usando Internet puede tener múltiples focos de interés, tales como la utilidad para varios perfiles de usuario: población general, personal médico o investigadores. Para dar soporte a los tres tipos de usuarios, los desarrollos emprendidos por el National Institutes of Health (NIH[2]), parte del U.S. Department of Health & Human Services, son dignos de mención. Dos ejemplos de los recursos disponibles por la población son: Healthfinder y MedlinePlus.

En nuestro caso algunos de estos recursos son tomados como referencias, particularmente los siguientes: (1) PubMed, que es usado para obtener casos científicos. Es una de las fuentes más utilizadas para buscar literatura biomédica (1300 millones de búsquedas en 2009). (2) La base de datos MIMIC II (Clifford G. D. et al, 2010) y los historiales clínicos de FMOD han sido seleccionados para probar la lógica lingüística en textos médicos, del mismo modo que han sido usados en otros trabajos relacionados (Luo G. and Tang C., 2008).

Para realizar la valoración previa al sistema de evaluación de la herramienta, hemos utilizado historiales médicos de MIMIC-II junto con casos facilitados por Pathology Department de Pittsburgh University[3] (elegidos al azar los casos: 223, 410, 474, 564, 565 y 616). Para mostrar la aplicación de la herramienta disponemos de uno de estos informes de casos recientes.

Después de introducir el texto en la herramienta y procesarlo, el reconocimiento de entidades medicas y las fuentes se muestra en una tabla que permite organizar y mostrar ambos. Los resultados se pueden ver en la Figura 2. El mismo proceso se puede aplicar al texto completo del caso.

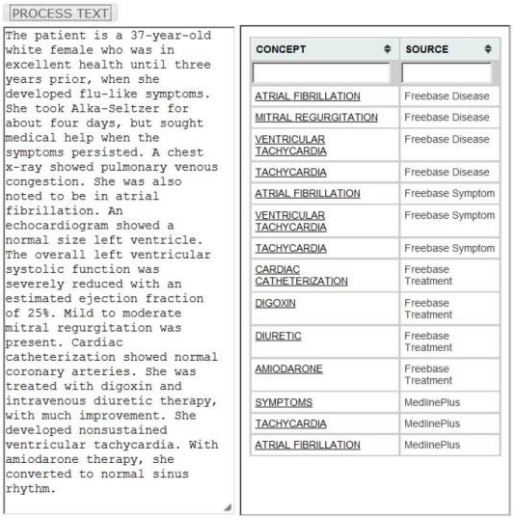

Figura 2: Resultados de la herramienta después del procesado de texto

Los conceptos resultantes están conectados a una página que muestra la información obtenida a través de los servicios ofrecidos por las fuentes. Por

[2] www.nih.gov

[3] path.upmc.edu/cases.html

ejemplo, el concepto "fibrilación auricular" se ha considerado como condición o síntoma en Freebase y también en Medlineplus (lo que indica que una entrada se realiza a través de la ontología NCBO llamada Medlineplus Health Topics).

Profundizando en los tópicos de enfermedades en Freebase, además de la descripción, podemos obtener un conjunto de síntomas, factores de riesgo y tratamientos asociados a una enfermedad en particular. Cada uno de estos términos relacionados semánticamente, están conectados a un tópico de la fuente. Si es necesario profundizar en los síntomas, la información semántica puede obtenerse a través de propiedades como "efectos secundarios" o "síntomas de". En el caso de buscar tratamientos, la información se obtiene a través de sus contraindicaciones, efectos secundarios y pruebas.

Si se selecciona Medlineplus como fuente, además de la descripción, se pueden obtener otros resultados relacionados con la investigación, como *Arrhythmia* o *Blood Thinners* entre otros, así como MeSH[4] (Medical Subject Heading) muestra sinónimos (los términos en MeSH para Arrhythmia son Arrhythmias y Cardiac).

Por último, tenemos la posibilidad de capturar una lista de publicaciones científicas en PubMed[4] utilizando la información obtenida en la búsqueda en Freebase y Medlineplus. En el primero, relacionamos los conceptos de la búsqueda a publicaciones relacionadas, en la segunda, relacionamos con los sinónimos de MeSH.

3.2 Aplicación al Dominio Turístico

El tipo de dominio al que nos enfrentamos tiene ciertas características que lo hacen especialmente interesante como caso de estudio para el sistema automático de extracción de información.

Lo primero, el léxico turístico utiliza una amplia terminología, tomada de varias áreas (geografía, economía, historia del arte, etc.). Es, por tanto, un uso del lenguaje poco especializado, un vocabulario amplio y común aunque hay un núcleo más específico de vocabulario, incluyendo términos técnicos relacionados con organizaciones turísticas, servicios, etc (Muñoz G. R. 2011). El tipo de información específica que se soporta (videos de monumentos o destinos turísticos) se caracteriza principalmente por la presentación de una amplia variedad de información. El sistema procesa texto turístico-cultural extraído de estos videos, identificando términos relevantes contenidos en Freebase. El contexto del turismo está consagrado por la meta de recuperar y extraer información, con particular énfasis en el sistema de recuperación de información personalizada, enmarcado en el consorcio Mavir, donde el plan se ha llamado: "Mejorando el acceso, el análisis y la visibilidad de la información y los contenidos multilingüe y multimedia en red para la Comunidad de Madrid". Mavir (http://www.mavir.net)

Los videos han sido desarrollados por el Instituto Español de Turismo (TURESPAÑA) que es la organización dependiente de la Administración General para la promoción de España como destino turístico. (http://www.tourspain.es).

En el caso del turismo, para ilustrar el uso, podemos seleccionar videos en la Web[5] y ver los resultados del sistema después de procesar el texto englobados en las siguientes categorías:

- Atracción turística: Una atracción turística es un lugar o rasgo característico que se visitaría como turista. Los ejemplos incluyen monumentos, parques, museos y similares. P ej.: Museo de Historia Natural de Beijing, Disneyland Park…
- Alojamiento: Este tipo está pensado para hoteles, bed and breakfast, hostales o cualquier otro lugar donde puedes quedarte al viajar, ej.: Hotel Palace San Francisco.
- Destino de viaje: un destino de viaje es un lugar donde se va en vacaciones, ej.: Paris o Bali.

El procesado del texto junto con los resultados de los conceptos identificándose muestra junto con otras dos maneras de conseguir información: (a) relacionándola directamente con la web de Freebase o (b) recuperando información desde el sistema directamente. (Ver Fig 3).

[5] www.ncbi.nlm.nih.gov/pubmed

[6] http://orion.esi.uem.es:8080/TouristFace/

Figura 3: Resultados del Sistema después de procesar el texto del video turístico

4 *Conclusiones y trabajo futuro*

En este artículo, hemos presentado un sistema que se caracteriza por la incorporación de las entidades anteriormente mencionadas procedentes de Freebase, para así reconocer conceptos médicos y turísticos en sistemas de entrada de videos y textos, de este modo ayudar a los usuarios a aumentar la cantidad de información obtenida. Asimismo se ha mostrado la manera en que se relacionan los términos identificados con Freebase. Además, hemos descrito la arquitectura de la plataforma y el desarrollo sistemático de un método para optimizar la integración de nuevas funcionalidades basadas en GATE. Por último, se ha ejemplificado el manejo del sistema.

En futuros trabajos incluiremos una evaluación sistemática por varios grupos de usuarios así como nuevas estrategias de lógica computacional, la incorporación de otras fuentes y nuevas interfaces de usuario, representación de conceptos mediante webs semánticas y la integración de otros sistemas en la arquitectura común escalable desarrollada.

Además, dentro de la plataforma de integración de información, se incluirán funciones de recomendación y personalización y las herramientas necesarias para el paso de medicina traslacional a medicina personalizada.

En el caso del dominio Turismo, se incluirá un módulo para la extracción automática de textos desde video ya desarrollado en el consorcio Mavir y la incorporación de alguna de las ontologías mencionadas.

Hasta donde alcanza la evaluación, se ha desarrollado un test de comprensión de casos clínicos a estudiantes de segundo año de Medicina, para medir la utilidad de una forma objetiva y subjetiva. El resultado del test será publicado próximamente.

En relación a los recursos, estamos estudiando la posibilidad de integrar nuevos contenidos procedentes de fuentes ya en uso (nuevas listas de Freebase y otras ontologías o recursos NCBO) junto con nuevas relaciones entre los resultados. Finalmente, con respecto a las nuevas interfaces, estamos estudiando el acceso a la herramienta desde dispositivos móviles.

Agradecimientos

Esta investigación ha sido financiada por el Ministerio de Ciencia y Tecnología Español MEDICAL-MINER (TIN-2009-14057-C03-01) y por la Comunidad de Madrid bajo el auspicio de la red de investigación MA2VICMR (S2009/TIC-1542)

Referencias

Allan J., B. Carterette y J. Lewis. 2005. When will information retrieval be “good enough”? SIGIR Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval

Aparicio, F., M. De Buenaga, M. Rubio, y A. Hernando. 2011a. An Intelligent Information Access system assisting a Case Based Learning methodology evaluated in higher education with medical students. *Computers & Education.*

Aparicio, F., R. Muñoz, M. Buenaga, y E. Puertas. 2011b. MDFaces: An intelligent system to recognize significant terms in texts from different domains using Freebase. En *Procesamiento de Lenguaje Natural*, 47, pp. 317-318.

Bollacker, K., C. Evans, P. Paritosh, T. Sturge y J. Taylor. 2008. Freebase: a

collaboratively created graph database for structuring human knowledge. In SIGMOD '08: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data held* in Vancouver, Canada, 1247-1250. ACM.

Brenes, D. J., D. G. Avello y K. P. González. 2009. Survey and evaluation of query intent detection methods. In WSCD '09: Proceedings of the 2009 workshop on Web Search Click Data held in Barcelona, Spain, 1-7. ACM.

Carrero, F., J. M. Gómez., M. de Buenaga, J. Mata, y M. Maña. 2007. Acceso a la información bilingüe utilizando ontologías específicas del dominio biomédico. *Procesamiento de Lenguaje Natural*, Vol. 38, pp. 107-117.

Chang, C., H. Kayed, M. Girgis y R. Shaalan. 2006. A Survey of Web Information Extraction Systems. IEEE Transactions on knowledge and data engineering. Volume: 18, Issue: 10

Clark, D. B., S. Touchman, M. Martinez-Garza, F. Ramirez-Marin, y T. Skjerping Drews. 2012. Bilingual language supports in online science inquiry environments. *Computers & Education*, 58(4), pp. 1207-1224.

Clifford, G., D. J. Scott y M. Villarroel. 2010. User Guide and Documentation for the MIMIC II Database, Rev: 259. Cambridge, MA, USA.

Cunningham H. et al, 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia.

De Melo, G., y G. Weikum. 2010. MENTA: inducing multilingual taxonomies from wikipedia. En *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 1099–1108, New York (USA)

Egozi, O., Markovitch S. y Gabrilovich E. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis.ACM Transactions on Information systems. Vol. 29 Issue 2

Gutiérrez, Y., A. Fernández, A. Montoyo y S. Vázquez. 2010. Integración de recursos semánticos basados en WordNet. Sociedad Española para el Procesamiento del Lenguaje Natural. Revista 45.

Hamburg, M. A. y Collins F. S. The Path to Personalized Medicine. New England Journal Med 2010; 363:301-304

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.

Knoth, P., T. Collins, E. Sklavounou, y Z. Zdrahal. 2010. Facilitating cross-language retrieval and machine translation by multilingual domain ontologies. En *Workshop on Supporting eLearning with Language Resources and Semantic Data* (at LREC 2010), Valletta (Malta).

Lew, M. S., N. Sebe, C. Djeraba y R. Jain Content-based multimedia information retrieval: state of the art and challenges. 2006. Journal ACM Transactions on Multimedia computing, Communications and Applications. Vol.2 Issue 1

Luo, G. y C. Tang. 2008. On iterative intelligent medical search. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval held in Singapore, Singapore, 3-10. ACM.

Mayfield, J., D. Lawrie, P. McNamee, y D. W. Oard. 2011. Building a Cross-Language Entity Linking Collection in Twenty-One Languages. En *P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. Rijke (Eds.), Multilingual and Multimodal Information Access Evaluation*, Vol. 6941, pp. 3-13, Berlin (Heidelberg).

Müller, C., y I. Gurevych. 2009. Using Wikipedia and Wiktionary in domain-specific information retrieval. En

Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access, CLEF'08, pp. 219–226, Berlin (Heidelberg)

Muñoz, R., F. Aparicio, M. De Buenaga. 2011. Tourist Face: A contents system base on concepts of freebase for Access to the cultural-tourist information. NLDB.

Nadeau, D. y Sekine S., 2007. A survey of named entity recognition and classification. En *Linguisticae Investigationes*, Vol. 30, pp. 3–26.

Navigil R. y P. Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites.

Tuchinda, R., C. Knoblock, A. y P. Szekely. Building Mashups by Demonstration. 2011. *ACM Transactions on the Web* (TWEB)

Voss, A., K. Nakata y M. Juhnke. Concept indexing. 1999. *Proceedings of the internation ACM SIGGROUP conference on Supporting group work.*

Verdejo, F., J. Gonzalo, D. Fernández, A. Peñas y F. López. 2000. ITEM: un motor de búsqueda multilingüe basado en indexación semántica. *Proceedings JBIDI.*