

Acceso a la información bilingüe utilizando ontologías específicas del dominio biomédico

Francisco Carrero García
José María Gómez Hidalgo
Manuel de Buenaga Rodríguez
Universidad Europea de Madrid
28035 Villaviciosa de Odón, Madrid, SPAIN {francisco.carrero,jmgomez,buenaga}@uem.es

Jacinto Mata
Manuel Maña López
Universidad de Huelva
Escuela Politécnica Superior
21071 Palos de la Frontera, Huelva, España manuel.mana@diesia.uhu.es, mata@uhu.es

Resumen: Unos de los enfoques más prometedores en la Recuperación de Información Croslingüe es la utilización de recursos léxico-semánticos para realizar una indexación conceptual de los documentos y consultas. Hemos seguido esta aproximación para proponer un sistema de acceso a la información para profesionales sanitarios, que facilita la preparación de casos clínicos, y la realización de estudios e investigaciones. En nuestra propuesta se conecta la documentación de los pacientes (la historia clínica), en castellano, con la información científica relacionada (artículos científicos), en inglés y castellano, usando para ellos recursos de gran cobertura y calidad como la ontología SNOMED. Se describe asimismo como se gestiona la confidencialidad de la información.

Palabras clave: Recuperación de Información Croslingüe, información biomédica, ontologías, recursos léxicos y semánticos, Unified Medical Language System (UMLS), SNOMED, Medical Subject Headings (MeSH)

Abstract: One of the most promising approaches to Cross-Language Information Retrieval is the utilization of lexical-semantic resources for concept-indexing documents and queries. We have followed this approach in a proposal of an Information Access system designed for medicine professionals, aiming at easing the preparation of clinical cases, and the development of studies and research. In our proposal, the clinical record information, in Spanish, is connected to related scientific information (research papers), in English and Spanish, by using high quality and coverage resources like the SNOMED ontology. We also describe how we have addressed information privacy.

Keywords: Cross-Language Information Retrieval, biomedicine, ontologies, lexical and semantic resources, Unified Medical Language System (UMLS), SNOMED, Medical Subject Headings (MeSH)

1 Introducción

La posibilidad de acceder, utilizando diferentes medios y desde cualquier lugar, al gran volumen de información que se genera cada día en el mundo es el elemento que caracteriza, cada vez más, la época actual. En este marco de innumerables ventajas, también cobra un peso creciente el problema general de la sobrecarga de información, y se hace cada vez mayor la necesidad del desarrollo de técnicas que ayuden a los usuarios a organizar, buscar y comprender la información (Buenaga, Fernández-Manjón y Fernández-Valmayor, 1995).

En esta situación, general, se encuentra también, con especial relevancia, el ámbito médico: los investigadores y profesionales en general de este ámbito, necesitan de forma cada vez más crucial, herramientas que faciliten el acceso a la información adecuada a sus necesidades (Hersh y Bhupatiraju, 2003).

Como agravante de la situación descrita, la información se encuentra disponible en múltiples idiomas, y no siempre la más relevante se encuentra disponible en el idioma materno del usuario, lo que no impide que sea comprensible para el mismo. Es necesario superar las barreras del lenguaje para entregar al usuario información en varios idiomas, ante

consultas suyas en uno solo. Ya no se trata de un entorno multilingüe, sino de recuperación *crosslingüe* – *Cross-Language Information Retrieval* ó CLIR (Grefenstette, 1998). Un entorno de trabajo tan retador exige soluciones nuevas, que pasan por la utilización creciente de recursos léxico-semánticos o de sistemas de traducción de gran cobertura y calidad.

En este artículo presentamos una propuesta de método de acceso a la información para el profesional sanitario, que se basa en asociar distintos tipos de información (especialmente clínica y científica) en dos idiomas. El modo previsto de trabajo es la presentación de información científica en inglés y castellano, relacionada de manera conceptual con la historia clínica del paciente objetivo. Esta propuesta se basa en la utilización de una ontología multilingüe específica del dominio biomédico para la representación de los documentos textuales, concretamente SNOMED (Spackman, Campbell y Côté, 1997). La asociación de conceptos de SNOMED a los documentos objetivo se aborda como una tarea de categorización automática (Sebastiani, 2002), y la asociación entre documentos de varios idiomas emplea el Modelo del Espacio Vectorial (Salton, 1989) usando como vocabulario de indexación los conceptos de la ontología.

El trabajo aquí descrito se encuentra enmarcado dentro de los proyectos de investigación SINAMED e ISIS¹ (Maña et al., 2006), cuyo objetivo es desarrollar nuevos mecanismos de acceso a la información mediante la aplicación de técnicas de análisis del lenguaje humano, en el ámbito de la biomedicina.

Hemos organizado este trabajo del modo siguiente. En la sección 2 se presenta el esquema general de acceso bilingüe, y se presenta brevemente el aspecto funcional de

nuestra propuesta. En la siguiente sección se describen y comparan los recursos léxico-semánticos más adecuados para nuestro trabajo. En la sección 4 se introducen las fuentes de información utilizadas en nuestro enfoque, junto con las técnicas actuales y las que nosotros hemos empleado hasta el momento para tratar la información confidencial. A continuación presentamos los elementos técnicos más relevantes de nuestra propuesta, finalizando este trabajo con una presentación de nuestros siguientes pasos.

2 Acceso bilingüe a la información biomédica

2.1 La Recuperación de Información Croslingüe

La globalización de la información, especialmente a través de Internet, exige que los sistemas de Recuperación de Información sean capaces de trabajar en entornos multilingües. Un entorno multilingüe es aquél en el que el usuario puede trabajar en varios idiomas, tanto a la hora de plantear consultas como a la de examinar resultados. Por ejemplo, algunos buscadores Web como Google, permiten la recuperación de resultados en múltiples idiomas ante una consulta en español.

Hay que resaltar que esta recuperación se realiza identificando los documentos en los que aparecen los términos de la consulta, independientemente del idioma de los documentos. Por ejemplo, la consulta “Java” podría arrojar resultados en múltiples idiomas, al tratarse de un nombre propio. Sin embargo, la consulta “lenguajes de programación orientados a objetos” difícilmente arrojaría resultados que no fuesen en castellano. De manera adicional, los resultados se pueden traducir al idioma nativo del usuario, usando sistemas de traducción automática.

Obviamente, este tipo de sistemas ofrecen funcionalidades multilingües limitadas. En los últimos años hemos asistido a un creciente interés por parte de investigadores y desarrolladores en los sistemas de Recuperación de Información Croslingüe – CLIR (Grefenstette, 1998). En este tipo de sistemas, se ofrece la posibilidad de superar de una manera efectiva las barreras del idioma, recuperando documentos en múltiples idiomas ante consultas en uno solo, de manera eficaz. Siguiendo el ejemplo anterior, un sistema croslingüe recuperaría documentos en

¹ SINAMED (Diseño e integración de técnicas de generación de resúmenes y categorización automática de textos para el acceso a información bilingüe en el ámbito biomédico) está parcialmente financiado por el Ministerio de Educación y Ciencia (TIN2005-08988-C02-01). ISIS (Sistema Inteligente de Acceso Integrado a la Información de Historial Clínico del Paciente y Documentación Médica Relacionada), ha sido parcialmente financiado por el Ministerio de Industria (FIT-350200-2005-16).

Este trabajo ha contado también con la financiación de la Comunidad Autónoma de Madrid, bajo la red de I+D MAVIR (S-0505/TIC-0267)

castellano e inglés ante la consulta “lenguajes de programación orientados a objetos”, con la misma efectividad que si la consulta también se hubiese expresado en inglés, como “*object-oriented programming languages*”.

Si Internet constituye un marco de referencia para los sistemas de recuperación croslingüe, debido a la abundancia de información en una gran cantidad de idiomas (por ejemplo, Wikipedia), no menos lo es el dominio de la biomedicina. No sólo recursos como MedLine indexan y ofrecen el acceso a medio millón de nuevas referencias al año², sino que los médicos se ven enfrentados de una manera diaria a la tarea de preparar casos de pacientes en base a información científica frecuentemente en otros idiomas. Si la necesidad de sistemas de recuperación croslingüe se hace patente al examinar la búsqueda en la Web, con más razón existe en dominios como el de la biomedicina. En la próxima sección presentamos el esquema de una aplicación de acceso a la información bilingüe (inglés-castellano) para el dominio de la biomedicina, con múltiples aplicaciones para médicos, investigadores y estudiantes.

2.2 Una propuesta de sistema Bilingüe de Acceso a la Información

Nuestra experiencia en los proyectos SINAMED e ISIS, incluye la observación de las fuentes de información que utilizan los médicos en su trabajo diario, en entornos como el Hospital de Fuenlabrada. También hemos trabajado con investigadores biomédicos, y con estudiantes de distintas disciplinas médicas (fisioterapia, enfermería, etc.) en el marco docente de la Universidad Europea de Madrid. De dichas observaciones se desprende que los médicos, científicos y estudiantes trabajan con información cuando menos bilingüe, a la hora de preparar casos, o elaborar informes y trabajos técnicos.

Con el fin de proporcionar a estos usuarios un acceso más sofisticado y efectivo a la información relevante para su trabajo, hemos ideado un sistema de acceso a la información bilingüe que permite relacionar el documento base de trabajo, el historial clínico, con la información científica relevante al mismo. En este sistema, se presenta un documento principal de trabajo (típicamente la historia

² Según los “Key MEDLINE® Indicators” (NLMA, 2007), se han agregado, por ejemplo, 606.000 referencias en 2005, y 623.089 en 2006.

clínica de un paciente objetivo, en español), y se permite acceder a información científica relacionada con el mismo (usualmente, informes científicos aparecidos en revistas de biomedicina, en inglés y castellano). Nuestro sistema tiene tres tipos posibles de usuarios:

- Los médicos en ejercicio, al preparar un caso clínico de un paciente concreto.
- Los investigadores cuando están analizando un caso arquetípico.
- Los estudiantes de ciencias biomédicas cuando están preparando un caso teórico.

En los tres casos, el usuario precisa acceder a la información científica más relevante para el diagnóstico y la toma de decisiones sobre pruebas o tratamientos del paciente, bien de manera teórica o práctica. En los últimos dos casos, el historial clínico se ha de presentar convenientemente anonimizado³, para evitar que el usuario tenga acceso a datos protegidos por las leyes vigentes de protección de información⁴. El tema de la anonimización, sin ser el centro de este trabajo particular, se discute con detalle en la sección 5.

La información científica mencionada puede encontrarse en múltiples idiomas⁵. El problema tipo para un potencial usuario de nuestro sistema es encontrar información científica en inglés y castellano, en relación con un historial en castellano. El elemento clave de nuestra propuesta es el modo de realizar esta conexión entre documentación médica en castellano y en otros idiomas, que tratamos en las próximas secciones.

2.3 Técnicas de Recuperación Croslingüe

Con el fin de enmarcar adecuadamente nuestro trabajo, se hace necesario discutir aunque sea

³ La anonimización es el proceso por el cual se eliminan o sustituyen todos los datos de un archivo de manera que no sea posible, en ningún caso, reconstruir la información original, identificado directa o indirectamente al sujeto o sujetos mencionados.

⁴ En el caso de la legislación nacional, el precepto más relevante es la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (frecuentemente designada como la LOPD).

⁵ Por ejemplo, en el portal PubMed (NLMb, 2007), se proporciona acceso a más de 33.000 revistas científicas en 60 idiomas, en muchos casos con el texto de los artículos parcial o totalmente disponible en la Web.

brevemente los enfoques generales para la Recuperación de Información Croslingüe. Existen múltiples taxonomías de métodos de este tipo, usualmente organizados en términos de los recursos utilizados para la tarea: diccionarios bilingües, córpora paralelos o comparables, etc. (Eichmann, Ruiz y Srinivasan, 1998; Gonzalo et al. 1998; Schauble y Sheridan, 1997; Volk et al. 2003). Por conveniencia, nosotros tomamos como referencia la clasificación realizada en el capítulo 4 de (Grossman y Frieder, 2004). En dicho capítulo se presentan, tres enfoques básicos para la recuperación croslingüe:

1. **Traducción de consultas.** La consulta se traduce a todos los idiomas objetivo y se recupera independientemente en cada uno de ellos, fusionando los resultados en una sola lista. Para ello, se utilizan recursos léxicos multilingües (diccionarios bilingües, tesauros, listas bilingües de términos generadas automáticamente, etc.).
2. **Traducción de documentos.** De manera alternativa a la anterior, se traducen todos los documentos al idioma de la consulta y se recupera en dicho idioma. Una ventaja importante sobre la traducción de consultas es que se dispone de más texto que en la consulta, y es de esperar que la traducción sea más precisa.
3. **Utilización de una representación interna para consulta y documentos.** En este enfoque, los documentos y la consulta se representan de una manera conceptual, típicamente independiente del idioma. El vocabulario de representación no está formado ya por los términos de los documentos, sino por los conceptos independientes del idioma a los que dichos términos hacen referencia. Identificados los conceptos a los que hace referencia una consulta, se recuperan los documentos indexados bajo ellos, independientemente de su idioma.

Los dos primeros métodos, no estando exentos de problemas que se escapan del ámbito de este artículo, son con diferencia los dominantes⁶. El tercer método claramente adolece de dos problemas:

- La práctica inexistencia de recursos lexico-semánticos de suficiente cobertura y calidad para un entorno genérico de recuperación, e.g. los motores de búsqueda en la Web.
- Las limitaciones en la eficacia de los sistemas de desambiguación de términos a significados y conceptos, que es también uno de los principales problemas en la calidad de los sistemas de traducción automática.

Sin embargo, cada vez existen más recursos del tipo requerido (aunque, desde luego, no de la cobertura necesaria), siendo quizá uno de los más representativos la base de datos léxica EuroWordNet (Gonzalo et al., 1998). Justamente en este trabajo se presenta cómo utilizar la componente independiente del idioma de EuroWordNet, el InterLingual Index (ILI), para realizar una recuperación croslingüe, de la manera más inmediata: usando los conceptos del ILI como elementos de indexación o índices. Sin embargo, la limitada cobertura no inglesa de EuroWordNet, junto con la baja efectividad de la desambiguación necesaria para identificar el concepto asociado a cada término, dificultan enormemente su implantación práctica y efectiva.

Sin embargo, existe un número creciente de recursos eventualmente utilizables en Recuperación de Información Croslingüe en biomedicina. En la próxima sección discutimos algunos de estos recursos y cómo se están utilizando. Por otra parte, el tipo de información con la que trabajamos en nuestra propuesta, permiten abordar los problemas de desambiguación con perspectivas de alcanzar la efectividad necesaria para que la recuperación sea precisa. Discutimos estas fuentes de información en la sección 4, para describir con más detalle nuestra propuesta en la 5.

3 Recursos biomédicos y uso en CLIR

En esta sección nos centramos en los tres recursos que, tras un análisis detallado conducido en las primeras fases del proyecto SINAMED, hemos considerado que se tratan de los tres de los más relevantes y utilizados en recuperación de información, especialmente croslingüe. Estos recursos son SNOMED, los MeSH y el UMLS.

⁶ En los artículos y libros citados previamente, el tercer enfoque prácticamente ni se menciona. Por otra parte, en la taxonomía de Grossman y Frieder (2004), este tercer enfoque está también basado en la

traducción, pero nosotros nos hemos permitido reinterpretarlo para acomodar nuestro razonamiento.

3.1 SNOMED-CT

SNOMED-CT (*Systematized Nomenclature of Medicine Clinical Terms*®) es una extensa terminología clínica desarrollada de manera conjunta por el *NHS Connecting for Health* y el *College of American Pathologists* (SNOMED Internacional, 2007).

La terminología SNOMED-CT cubre enfermedades, hallazgos clínicos y procedimientos, y ayuda a realizar indexación, almacenamiento, recuperación y agregación de datos clínicos de forma consistente. Para ello, permite estructurar y gestionar por ordenador los registros médicos, reduciendo la variabilidad en la manera en que se pueden adquirir, utilizar y codificar los datos necesarios para el cuidado clínico de los pacientes y la investigación.

Sus elementos básicos son:

- Conceptos: representan una unidad mínima de significado.
- Jerarquías: compuestas por categorías de primer nivel y sus correspondientes subcategorías.
- Relaciones: las de tipo “es_un” permiten enlazar conceptos con jerarquías; las relaciones de atributos conectan conceptos entre jerarquías.
- Descripciones: términos o nombres asociados a un concepto.

La última versión se compone de más de 308.000 conceptos organizados en 19 categorías jerárquicas de primer nivel. Además, contiene más de 770.000 descripciones y más de 924.000 relaciones.

Existe una versión en español de SNOMED-CT que mantiene el diseño técnico, la arquitectura, el contenido (tablas de conceptos, descripciones y relaciones, tablas de relaciones históricas, referencias cruzadas con la CIE, etc.), y la documentación técnica relacionada.

3.2 MESH

Los Medical Subject Headings (MeSH) son un tesoro desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos (NLMc, 2007). Se compone de una serie de términos asociados a descriptores, dispuestos en una estructura jerárquica que permite realizar búsquedas con diversos niveles de especificidad.

Los descriptores de MeSH se organizan de dos maneras distintas: la primera es una lista

alfabética de descriptores con las respectivas referencias cruzadas de sinónimos y términos relacionados; la segunda es una clasificación jerárquica que agrupa a todos los descriptores en 16 categorías, que se subdividen a su vez en subcategorías con un mayor nivel de especificidad.

Estos árboles de descriptores no constituyen una clasificación exhaustiva de las materias, y se utilizan como guía para las personas encargadas de asignar categorías a documentos.

En su última versión, MeSH dispone de 22.997 descriptores, así como de más de 151.000 conceptos suplementarios (*Supplementary Concept Records*) recogidos en un tesoro separado. Existen también más de 136.000 referencias cruzadas que ayudan a determinar el descriptor de MeSH más apropiado para cada caso.

La principal aplicación de MeSH se encuentra en su uso por parte del NLM para indexar artículos de más de 4.800 de las principales revistas biomédicas para la base de datos MEDLINE/PubMED (NLMa, 2007).

3.3 UMLS

El UMLS (*Unified Medical Language System*) es un sistema desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos. Está compuesto por un meta-tesoro, una red semántica y un lexicón especializado, distribuidos con una serie de herramientas que facilitan su uso (NLMd, 2007).

El meta-tesoro es una base de datos multilingüe y multipropósito que contiene información sobre conceptos biomédicos y relacionados con la salud, incluyendo sus diferentes nombres y sus relaciones.

La red semántica proporciona una clasificación consistente de todos los conceptos representados en el meta-tesoro, además de un conjunto de relaciones entre dichos conceptos. Todos los conceptos del meta-tesoro tienen asignado al menos un tipo semántico de la red semántica.

El lexicón especializado pretende ser un lexicón general que incluye términos biomédicos. La mayoría de los términos que aparecen en los nombres de conceptos del meta-tesoro aparecen igualmente en el lexicón.

Es de reseñar que UMLS se nutre de múltiples lexicones y ontologías, entre los que se encuentran tanto MeSH como SNOMED. De alguna manera, el UMLS es un “super-sistema”

que incluye a los anteriores y proporciona una estructura común a estos y otros recursos.

3.4 Utilización en CLIR

Los recursos léxico-semánticos anteriores han sido concebidos desde un principio con el fin de proporcionar modos de acceso más unificados y efectivos a la información biomédica. En conjunto, se han convertido en los vocabularios controlados de indexación de la información biomédica, permitiendo a los usuarios no sólo búsquedas con texto libre a diversos canales de información, sino también búsquedas conceptuales que han demostrado su efectividad en la práctica⁷ (Lowe y Barnett, 1994).

Con el fin de sistematizar toda la literatura médica, y no sólo la inglesa, han ido apareciendo versiones de los recursos en distintos idiomas, y de manera natural, se han empleado en el desarrollo de sistemas e investigaciones multilingües y croslingües. Por su relación con nuestro trabajo, destacamos los siguientes:

- En (Hersh y Donohoe, 1998) se presenta el sistema SHAPIRE Internacional, una adaptación del sistema de recuperación conceptual SHAPIRE a entornos multilingües. Este sistema permite recuperar conceptos de UMLS en inglés ante consultas en múltiples idiomas, incluyendo el castellano y el alemán. Al no devolver textos, no se puede hablar de una herramienta de recuperación de textos plena, pero si incluye su componente fundamental, que es el acceso a los conceptos independientes del idioma a partir de textos (consultas) en múltiples idiomas.
- En (Volk et. al, 2002) se describe el enfoque de indexación conceptual usando UMLS que se realiza en el marco del proyecto de investigación europeo MUCHMORE, con el fin de evaluar la viabilidad técnica de construir sistemas de CLIR basados en indexación conceptual interlingüe. El énfasis es en el nivel de procesamiento del lenguaje necesario para alcanzar niveles razonables de calidad en la indexación, que los experimentos permiten afirmar que son suficientes.

⁷ Una búsqueda en PubMed por “*UMLS and information and retrieval*” devuelve más de 200 resultados, correspondientes a informes científicos en los que UMLS se utiliza de alguna forma en un sistema de Recuperación de Información.

- En (Marko, Schulz y Hahn, 2005) se presenta el sistema MorphoSaurus, que realiza recuperación croslingüe usando UMLS para la indexación interlingüe, y se realiza un experimento que compara la efectividad de dicho enfoque con uno basado en traducción de consultas, resultando favorable al primero la evaluación.

Estos informes, junto con la naturaleza específica de la información con la que trabaja nuestro sistema (y que discutimos a continuación), nos permite concluir que nuestro enfoque es viable y muy prometedor en términos de efectividad.

4 Fuentes de información

La información médica es voluminosa y de extrema complejidad. Uno de los factores con una mayor repercusión en la heterogeneidad de los contenidos médicos es la diversidad de fuentes. Cada fuente (escritos científicos, bases de datos de resúmenes, bases de datos estructuradas o semi-estructuradas, servicios Web o historiales clínicos de pacientes) tiene diferentes elementos y aspectos, como por ejemplo, la existencia o no de una estructura externa del documento, la existencia de texto libre con datos estructurados (tablas con resultados clínicos) o la longitud de los documentos. Estas diferencias en dominio, estructura y escala, dificultan el desarrollo de sistemas robustos e independientes que faciliten el acceso a este tipo de contenidos. Esta dificultad se agrava con la naturaleza multilingüe de la información, y es a lo que pretendemos dar respuesta con nuestra propuesta.

En nuestra propuesta, se conectan dos tipos de información que se discuten a continuación. Dado que las historias clínicas contienen información sensible desde un punto de vista de la confidencialidad, también se discute su anonimización.

4.1 Documentación médica

Considerando por ejemplo, los artículos científicos médicos, hay miles de revistas científicas en inglés, y el problema crece si consideramos otros lenguajes y fuentes.

Medline, la base de datos bibliográfica más importante y consultada en el dominio biomédico constituye un ejemplo principal. Medline almacena referencias a artículos de revistas desde 1966 hasta la actualidad,

contiene más de 13 millones de referencias, con un crecimiento de entre 1.500 y 3.500 referencias por día. Esta gran cantidad de información hace difícil a los expertos sacar partido de toda la información publicada.

En los sistemas desarrollados en nuestros proyectos, para ser probados y evaluados sobre usuarios finales, y para el que nos ocupa en particular, hemos trabajado sobre conjuntos representativos de esta información. En concreto se ha seguido un criterio para seleccionar un conjunto de revistas considerando el lenguaje (castellano e inglés), relevancia de la revista al proyecto (estábamos especialmente interesados en neumonía, enfermedades del corazón y alumbramientos) y acceso libre al texto completo. Teniendo presentes estas guías se seleccionaron: *British Medical Journal*, *Journal of the American Association* y las revistas en castellano *Archivos de Bronconeumología* y *Anales de Pediatría*. Estas revistas publican artículos de diferentes clases, entre los que hemos seleccionado: *scientific papers* (trabajos de investigación originales), *clinical reviews* (revisiones de literatura disponible en un tema), *practice* (escritos breves que están centrados en historias de casos específicos), técnicas y procedimientos, y noticias.

4.2 Historiales clínicos

El historial clínico del paciente se define como el conjunto de documentos (datos, análisis, diagnósticos y otros tipos de información) que son generados a lo largo del proceso asistencial del paciente. El sistema de registros en papel clásico presenta toda una serie de limitaciones (información poco legible, desorganización, ausencia de consistencia, accesibilidad limitada, garantía incierta de confidencialidad, etc.) que pueden mejorarse con la utilización de registros electrónicos integrados.

Alguna de las ventajas del historial clínico electrónico son: mejor accesibilidad a la información y mejora en la confidencialidad, homogenización de datos, visión completa del paciente, coordinación de tratamientos médicos, etc.

En sistemas desarrollados en nuestros proyectos, hemos trabajado con información anonimizada en Español del hospital (Hospital de Fuenlabrada) que formaba parte del consorcio, de dos tipos: notas de evolución (9413 notas de evolución de 3666 historiales clínicos diferentes – una media 2,6 notas por

historial) e informes de alta (49 informes completos redactados al abandonar un paciente el hospital). En puntos siguientes se dan más detalles sobre este tipo de fuente en inglés.

4.3 Tratamiento de información confidencial

Los historiales clínicos almacenan información que puede ser de gran utilidad en la investigación médica. Sin embargo, como los historiales contienen también información confidencial estos deben ser tratados con la debida cautela. La Ley 16/2003 de Cohesión y Calidad del Sistema Nacional de Salud garantiza la confidencialidad e integridad de los datos en el intercambio de información entre los organismos del Sistema Nacional de Salud.

En general, el uso por terceros de información médica que incluya datos personales del paciente requiere el permiso expreso de este. Cuando la información que se desea tratar se encuentra almacenada de cierto tiempo, puede ser imposible conseguir este permiso. En este caso, la anonimización de la información clínica mantiene el nivel de confidencialidad deseado a la vez que permite el acceso a la información (Kalra et al., 2006).

La anonimización del historial médico de un paciente consiste en eliminar la información que puede identificar a las personas involucradas en el proceso asistencial; tanto el paciente como los profesionales sanitarios que lo atendieron. La legislación estadounidense, a diferencia de la europea, define en el *Health Information Portability and Accountability Act (HIPAA)* (US Government, 1996), los elementos que deben ser excluidos en el proceso de anonimización. De entre ellos, destacamos los que suelen aparecer en un historial clínico: nombre y apellidos de los pacientes, representantes legales y familiares; nombres y apellidos de los médicos; números de identificación; números de teléfonos, fax y buscapersonas; nombres de hospitales; direcciones y localizaciones geográficas; fechas. La relevancia, cada vez mayor, de esta tarea está estimulando la organización de congresos, talleres y competiciones como *i2b2: Challenges in NLP for Clinical Data: De-identification Challenge* (i2b2 NCBC, 2007).

Dentro del proyecto ISIS (Buenaga et al., 2006) se trabajó con dos tipos de documentos pertenecientes al historial clínico del paciente que, en el marco de dicho proyecto, fueron proporcionados por el Hospital de Fuenlabrada.

```

<PAC: "PACIENTE:">
<FPAC: ((("-")|("\n"))>
<FD: ("FDO")>
<DR: ("dr.")|("Dra")>
void exprBasica():>{ { <PAC>((nombre())(<BLANCO4>)?<FPAC>
|...
|<FD>firmado() }
void firmado():>{ {
(<BLANCO1>)*<DR>(<BLANCO1>)*(<DOSPUNTOS>)?nombre()(<FPAC>)?(<BLANCO3>)?(<BLANCO4>)?
|<DOSPUNTOS>(blancos())*(<DR>)?(<BLANCO1>)*(<BARRA><DR>)?(<PUNTO>)?(<BLANCO1>)*doctor()
|(<PUNTO>)?(<DOSPUNTOS>)?(<BLANCO1>)*(<DR>)?(<PUNTO>)? (<BLANCO1>)*fin_firmado()
}
}

```

Figura 1: Ejemplo de regla sintáctica para la eliminación de nombres de pacientes y médicos.

Estos tipos de documentos son las notas de evolución clínica y los informes de alta.

Las *notas de evolución clínica* son informes escritos por los médicos acerca de los cambios que se producen durante el proceso asistencial. El Hospital de Fuenlabrada proporcionó 9.413 notas de evolución pertenecientes a 3.666 historiales distintos, lo que supone una media de 2,6 notas por historial. El tratamiento de esta información requería un proceso previo de anonimización, ya que, contenían nombres de pacientes y médicos. La aproximación elegida, en este caso, fue la de analizar manualmente unas 100 notas de las que se extrajeron alrededor de 120 reglas sintácticas. Mediante la aplicación de estas reglas se consiguió eliminar, de forma automática, 393 nombres de médicos y pacientes. Finalmente, se eliminaron otros 30 nombres de forma manual.

En la Figura 1 se muestra un ejemplo de una de las reglas sintácticas utilizadas. Esta regla permite la identificación del nombre del paciente después de la palabra "PACIENTE:" o la del nombre del médico después de la aparición de la expresión "FDO Dr."

Los *informes clínicos de alta* constituyen un resumen del proceso asistencial del paciente que redactan los médicos al finalizar dicho proceso. El Hospital de Fuenlabrada proporcionó 49 informes de diferentes servicios hospitalarios: urgencias, urgencias pediátricas, cirugía general y digestiva, pediatría, maternidad, traumatología, medicina interna y medicina intensiva. Para la anonimización de estos informes se llevó a cabo un proceso similar al descrito para las notas de evolución. De esta manera, se eliminó cualquier información personal sobre datos de los pacientes y los médicos que los trataron.

A pesar de que la técnica empleada consigue anonimizar ambos tipos de documentos de forma efectiva, en la actualidad, la estrategia

más utilizada es la aplicación de aprendizaje automático. El problema de la anonimización se puede plantear como una tarea de reconocimiento de entidades nombradas (REN), donde las entidades que se desean identificar son los datos con carácter confidencial. Este es el enfoque seguido en todos los trabajos presentados al i2b2. Los participantes en la competición dispusieron de una colección de entrenamiento formada por 671 informes de altas escritos en inglés que incluyen 14.309 entidades con información de carácter personal. En [Guo et al., 2006] los autores hacen uso de Support Vector Machines sobre características a nivel de token y otras específicas para cada tipo de entidad a reconocer. Otra aproximación distinta es la que se propone [Aramaki et al., 2006], donde además de características locales emplean otras dos de carácter global: información de las frases previa y siguiente, y consistencia de etiquetas de clasificación en el informe y en el corpus. También, en una componente de nuestro proyecto SINAMED que hemos evaluado de forma preliminar sobre los datos de I2B2 (para Smoking Challenge), hemos conseguido unos resultados muy cercanos a la media construyendo el clasificador únicamente utilizando atributos léxicos y morfológicos, sobre la arquitectura que estamos desarrollando y que describimos en el punto siguiente (en concreto un valor para la f-measure de 0,765 frente a 0,795 de la media). El uso de estos atributos léxicos y morfológicos en combinación con los conceptos de Snomed podrían ayudarnos a obtener mejores resultados.

5 Enfoque técnico de nuestra propuesta

Nuestra propuesta está basada en los elementos anteriormente descritos. Se propone el desarrollo y evaluación de un sistema de acceso a la información para profesionales y

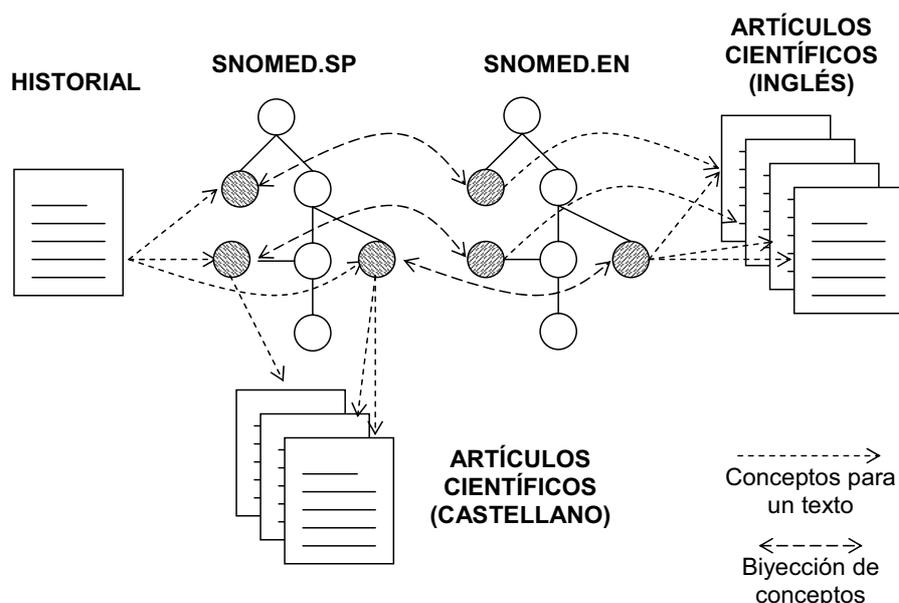


Figura 2: Método de conexión de del historial con la información científica relacionada vía e.g. SNOMED.

estudiantes sanitarios, en el que se relaciona la información básica de trabajo (la historia clínica), típicamente en castellano, con la información científica apropiada, típicamente en inglés y castellano. El objeto de este sistema es simplificar la preparación de casos, investigaciones o trabajos, al evitar la necesidad de realizar búsquedas explícitas de información científica, al tiempo que realizar esta búsqueda implícita con mayor eficacia que el propio usuario.

El esquema de asociación de las fuentes de información se presenta de forma gráfica en la figura 2. En dicha figura se observa como los historiales médicos se asocian a conceptos de e.g. la ontología de SNOMED en castellano, cuyos conceptos están a su vez asociados a los de SNOMED en inglés de una manera cuasi-biyectiva. Por medio de los conceptos en español, se recuperan documentos científicos en español. También se recuperan documentos científicos en inglés usando estos conceptos en inglés, que han sido también asociados de manera automática con dichos documentos.

Los pilares del sistema son:

- La utilización de técnicas de recuperación croslingüe basadas en indexación conceptual interlingüe, avalada por otros trabajos, y que en nuestro caso se simplifica al no tener que desambiguar consultas sino fragmentos de información más extensos (los historiales clínicos). Inicialmente, y en vista del interés demostrado por los médicos que actuarán

como usuarios del sistema, se está utilizando la ontología SNOMED.

- La utilización de técnicas de categorización automática (Sebastián, 2002), y no de desambiguación, para la asignación de conceptos de SNOMED a los documentos objetivo.

Nosotros entendemos que en gran medida, los conceptos de SNOMED y en general del UMLS son más categorías temáticas que conceptos semánticos de grano fino como los de e.g. EuroWordNet, por lo que se pretende evitar una aplicación término a concepto, y promover una sistema texto a categoría. Los sistemas de categorización basados en aprendizaje han alcanzado niveles de efectividad comparables a los de profesionales humanos entrenados. Nuestra experiencia en este sentido es prolongada⁸, y avala nuestras perspectivas.

Gran parte de la información médica científica se haya clasificada de acuerdo a los vocabularios conceptuales estándar mencionados anteriormente. Sin embargo, la información de los historiales médicos no está clasificada de esta manera. Esto supone un alimitación, dado que nos proponemos realizar la clasificación usando sistemas basados en aprendizaje, que dependen de la existencia de material manualmente clasificado para su

⁸ Véase como guía e.g. (Gómez et al., 2004; Gómez, Buenaga y Cortizo, 2005).

entrenamiento. Lo habitual en estas situaciones es utilizar una técnica de *bootstrapping*, que consiste en clasificar un conjunto semilla de documentos, usarlos para entrenar el sistema, clasificar con él un segundo grupo de documentos, y revisar manualmente las decisiones menos seguras. Repetido iterativamente, este proceso permite construir una colección de datos de una magnitud suficiente de manera efectiva. Una vez obtenida esta colección, el sistema se entrena sobre ella, alcanzando niveles de calidad adecuados en sus decisiones sobre nuevos documentos.

6 Conclusiones y trabajo futuro

En este artículo, se ha presentado una visión de cómo conseguir el acceso a informes científicos en inglés y castellano a partir de un historial en castellano, utilizando para ello una categorización automática respecto a una ontología bilingüe. También se han discutido las diferencias fundamentales entre dos de las ontologías más relevantes en el ámbito biomédico: SNOMED y MESH. Se han descrito las fuentes de información más significativas en el marco del problema, considerando el aspecto fundamental de la confidencialidad de la información médica que incluye datos de carácter personal. Para solventar este problema, se ha expuesto la solución utilizada sobre dos colecciones de documentos proporcionadas por el Hospital de Fuenlabrada y se han discutido soluciones distintas sobre colecciones de informes de alta en inglés.

En el futuro planificamos integrar en un sistema, la categorización de los documentos, la recuperación de los mismos y la anonimización de los informes médicos. Este sistema debe permitir un acceso personalizado en función del perfil del usuario. Se han concebido tres perfiles de usuario para el sistema: médicos en atención hospitalaria, investigadores médicos y alumnos de titulaciones relacionadas con la biomedicina. Con la ayuda de un número significativo de usuarios de cada perfil, se diseñarán las interfaces adecuadas.

Una vez completada esta primera fase, hemos planificado la realización de implementaciones más efectivas de los distintos clasificadores que forman el sistema. Estas nuevas implementaciones se evaluarán sobre colecciones de referencia, como la utilizada en i2b2 para la anonimización.

Finalmente, integraremos los clasificadores en la herramienta y se llevarán a cabo experimentos que permitan validar la utilidad del sistema con cada uno de estos perfiles.

Bibliografía

- Aramaki, E., Miyo, K. Automatic Deidentification by Using Sentence Features and Label Consistency. Proceedings of the Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
- Buenaga, M., Fernández-Manjón, B., Fernández-Valmayor, A, “Information Overload at the Information Age”. Collis, B., Davies, G. (eds) “Innovating Adult Learning with Innovative Technologies”, Ed. Elsevier, 1995.
- Buenaga, M., Maña, M.J., Gachet, D., Mata, J., 2006. The SINAMED and ISIS Projects: Applying Text Mining Techniques to Improve Access to a Medical Digital Library. LNCS: Research and Advanced Technology for Digital Libraries, vol. 4172, pp. 548-551.
- Eichmann, D., Ruiz, M.E., y Srinivasan, P. , 1998. Cross-Language Information Retrieval with the UMLS Metathesaurus. SIGIR'98 - 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24 – 28.
- Gómez, J.M., Cortizo, J.C., Puertas, E., Ruíz, M., 2004. Concept Indexing for Automated Text Categorization. In Natural Language Processing and Information Systems: 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004, Salford, UK, June 23-25, 2004, Proceedings, Lecture Notes in Computer Science, Vol. 3136, Springer, pp. 195-206.
- Gómez, J.M., Buenaga, M. de, Cortizo, J.C., 2005. The Role of Word Sense Disambiguation in Automated Text Categorization. Montoyo, A.; Muñoz, R.; Métais, Elisabeth (Eds.), Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, Proceedings, Lecture Notes in Computer Science, Vol. 3513, Springer, pp. 298-309.

- Gonzalo, J., Verdejo, F., Peters, C. y Calzolari, N., 1998. Applying EuroWordNet to Cross-Language Text Retrieval. *Computers and the Humanities*, 32, 2-3, 185-207.
- Grefenstette, G., (ed.) 1998. Cross-language information retrieval. The Kluwer international series on information retrieval 2, Kluwer Academic.
- Grossman, D.A., Frieder, O., 2004. *Information Retrieval: Algorithms and Heuristics*. Second Edition. Springer.
- Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., Hepple, M., 2006. Identifying Personal Health Information Using Support Vector Machines. *Proceedings of the Workshop on Challenges in Natural Language Processing for Clinical Data*.
- Hersh, W.R., Donohoe L.C., SAPHIRE International: a tool for cross-language information retrieval. *Proceedings of the 1998 AMIA Annual Symposium*, 1998, 673-677.
- Hersh, W. y Bhupatiraju, R.T., 2003. TREC Genomics Track Overview. *NIST Special Publication: SP 500-255 (The Twelfth Text Retrieval Conference)*, pp. 14-23.
- i2b2 (Informatics for Integrating Biology and the Bedside) National Center for Biomedical Computing (NCBC), 2007. *Challenges in Natural Language Processing for Clinical Data*. URL: <https://www.i2b2.org/NLP/>. Acceso: 28 de enero de 2007.
- Kalra, D., Gertz, R., Singleton, P., Inskip, H.M., 2006. Confidentiality of personal health information used for research. *British Medical Journal*, vol. 333, pp. 196-198.
- Lowe, H. y Barnett, G. 1994. Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *Journal of the American Medical Association*, 271(14):1103-1108.
- Marko, K., Schulz, S., Hahn, U., 2005. MorphoSaurus--design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4), pp. 537-45.
- NLM (National Library of Medicine), 2007. Key MEDLINE® Indicators. URL: http://www.nlm.nih.gov/bsd/bsd_key.html. Acceso: 28 de enero de 2007.
- NLM (National Library of Medicine), 2007. PubMed. URL: <http://www.pubmed.gov/>. Acceso: 28 de enero de 2007.
- NLM (National Library of Medicine), 2007. Medical Subject Headings. URL: <http://www.nlm.nih.gov/mesh/>. Acceso: 28 de enero de 2007.
- NLM (National Library of Medicine), 2007. Unified Medical language System. URL: <http://www.nlm.nih.gov/research/umls/>. Acceso: 28 de enero de 2007.
- Salton, G. 1989. *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley, Reading, US.
- Schauble, P. y Sheridan, P., 1997. Cross-Language Information Retrieval (CLIR) Track Overview. *The Sixth Text REtrieval Conference (TREC-6)*, National Institute of Standards and Technology (NIST), Special Publication 500-240.
- Sebastiani, F. 2002. *Machine Learning in Automated Text Categorization*. *ACM Computing Surveys*, 34(1):1-47.
- SNOMED International, 2007. SNOMED-CT. URL: <http://www.snomed.org/snomedct>. Acceso: 28 de enero de 2007.
- Spackman, K.A., Campbell, K.E, Côté, R.A., 1997. SNOMED-RT: a reference terminology for health care. *Proceedings of the AMIA Annual Fall Symposium*, pp. 640-4.
- US Government, 1996. *Health Information Portability and Accountability Act*. Washington, D.C.: US Government Printing Office.
- Volk M, Ripplinger B, Vintar S, Buitelaar P, Raileanu D, Sacaleanu B., 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67 (1-3), pp. 97-112.