

Máster en Investigación Informática

Facultad de Informática  
Universidad Complutense de Madrid

Proyecto de Fin de Máster en Ingeniería de  
Computadores

---

FastUMLS: Extracción de conceptos en textos biomédicos

---

Autor: José Luis Marina Leonardo  
Director: Alberto Pascual Montano  
Departamento de Arquitectura de Computadores  
y Automática  
Facultad de Informática  
Universidad Complutense de Madrid

**Curso 2008/2009**



**Abstract:**

Identifying and extracting the key concepts from Biomedical texts is getting increasingly important because of the exponential growth of scientific literature, massive gene experiments, and diseases, drugs or process descriptions in the Medical and Molecular Biology fields. In this work we introduce FastUMLS, an automatic annotation tool to identify formal concepts from the *Unified Medical Language System* (UMLS) Ontology from biomedical texts. Instead of employing the traditional Natural Language Processing techniques the proposed method is based on concept scoring using different frequencies of the terms in the concept definition. Different experiments have been made in order to compare the results between the proposed approach and NLP tools, and data about the controlled revision of the resulting concepts by the authors of different scientific papers is as well presented. This preliminary results show that FastUMLS can effectively extract concepts from Biomedical texts within reasonable performance times.

FastUMLS is freely accessible at <http://bisaurin.dacya.ucm.es:8283/fastumls>

**Keywords:**

Text annotation, information extraction, biomedical ontologies, text mining, UMLS

**Resumen:**

Identificar y extraer cuáles son los conceptos clave relacionados con un texto es un proceso que toma cada mayor importancia en Medicina y en Biología Molecular dado el incremento exponencial de información existente en forma de artículos científicos, descripciones de experimentos, y anotaciones a enfermedades, medicamentos y procesos. En este trabajo se muestra una herramienta para la anotación automática de textos con conceptos formales definidos en la Ontología del *Metathesaurus* del *Unified Medical Language System* (UMLS). El método propuesto no se basa en el acercamiento tradicional propuesto por las herramientas de Procesamiento de Lenguaje Natural (PLN) sino en la ponderación de los conceptos de acuerdo a las frecuencias de aparición de las palabras buscadas en las definiciones. Se informa de los resultados de diferentes experimentos que comparan los resultados obtenidos con los conceptos ofrecidos por herramientas tradicionales y de pruebas revisadas por investigadores expertos y autores de artículos que han sido procesados por la herramienta. Los resultados demuestran que FastUMLS extrae de forma efectiva conceptos sobre textos biomédicos en tiempos más que razonables.

Puede accederse libremente a FastUMLS en la dirección:

<http://bisaurin.dacya.ucm.es:8283/fastumls>

**Palabras clave:**

Anotación de textos, extracción de información, ontologías biomédicas, minería de textos, UMLS

Como autor principal autorizo a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente trabajo Fin de Máster: FastUMLS: Extracción de conceptos en textos biomédicos, así como todo el código fuente y algoritmos utilizados, realizado durante el curso académico 2008-2009 bajo la dirección de Alberto Pascual-Montano en el Departamento de Arquitectura de Computadores y Automática, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Firmado:

José Luis Marina Leonardo

# Índice

Introducción.....	1
Ontologías.....	4
UMLS.....	7
Conceptos y Terminología.....	9
Conceptos e Identificadores de Concepto (CUIs).....	10
Términos e Identificadores Léxicos (LUIs).....	10
Nombres de Conceptos e Identificadores de Cadenas (SUI).....	10
Átomos e Identificadores de Átomos (AUI).....	10
Métodos.....	18
Objetivos.....	18
Arquitectura de Procesos.....	20
Preprocesado de los Datos.....	22
Procesamiento On-line.....	25
Resultados y Conclusiones.....	30
Tiempos y Capacidad de Cálculo.....	30
Pruebas Comparativas.....	31
Identificación de Conceptos en Genes Anotados por Expertos.....	31
Comparación de resultados con Herramienta de PLN.....	34
Discusión de los Resultados.....	36
Pruebas Supervisadas.....	37
Discusión de los resultados.....	41
Líneas Futuras.....	43
Identificación de Grupos de Conceptos .....	43
Enriquecimiento de Búsquedas:.....	43
Mejoras al proceso actual.....	44
Preprocesado de los datos y los pesos relativos:.....	44
Tratamiento Semántico de los datos de entrada:.....	44
Eliminación de Duplicados:.....	44
Software y herramientas.....	45
Referencias.....	47
Agradecimientos.....	50

## Introducción

En el presente trabajo que se sitúa en el ámbito de la Bioinformática, se expone una metodología para la extracción de conocimiento en forma de conceptos o términos, a partir de elementos de texto libre y en el ámbito de la Biomedicina. La propuesta es la de realizar esta tarea de forma automática y análoga a como hacemos los humanos cuando a partir de un resumen de una noticia podemos inferir los elementos principales e incluso resumirla en un título o en una serie de etiquetas.

En este caso los conceptos o ideas clave del texto analizado se corresponden con elementos preexistentes en una base de datos u ontología revisada por expertos en el campo de la Medicina y la Biología Molecular.

La Bioinformática es el uso de técnicas matemáticas e informáticas para almacenar, gestionar y analizar datos biológicos para responder a problemas de la biología [Kaminski 2000].

En el caso que nos ocupa nos centramos en textos Biomédicos y en lugar de utilizar pesadas técnicas de Procesamiento de Lenguaje Natural, el algoritmo se ha basado en permutaciones de las palabras claves del texto y sus conceptos relacionados en los vocabularios controlados y ontologías existentes en la red, seleccionando en un paso posterior aquéllos con mayor significancia estadística.

La idea es la de ofrecer a partir de un texto tipo - como un resumen de un artículo, anotaciones de un experimento o palabras claves de un estudio - qué conceptos - enfermedades, procesos, genes, proteínas, etc - son los que mejor definen el asunto y siendo estos conceptos elementos normalizados y no redundantes como son los de la base de datos unificada del *Unified Medical Language System*® (UMLS).

Dos son los problemas principales con que se topa una tarea como la expuesta:

- Primero necesitamos contar con una base de datos de conceptos sobre el área de interés y a los que hacer referencia una vez procesado un texto.
- Segundo, cómo resolver problemas como la clasificación o importancia relativa de los conceptos extraídos de forma que el resultado tenga sentido para un lector experimentado.

Otro tipo problemas, no menos complejos, son aquellos que atañen a la minería de textos como son la desambiguación de conceptos, la identificación de frases o grupos de palabras relacionadas o la identificación de sinónimos, por nombrar algunos de ellos.

La necesidad de estandarizar el vocabulario utilizado en diferentes campos de la ciencia no es nuevo. Poder utilizar el mismo término para referirse a determinado fenómeno o a una enfermedad, de forma única y sin ambigüedades es una necesidad que tratan de resolver las bases de datos de anotaciones y las ontologías.

En el caso de la Biología Molecular y de la Bioinformática esta necesidad se manifiesta cada vez más urgente dada la explosión de datos de las últimas décadas. Son éstas disciplinas que se han desarrollado de forma paralela a la acumulación de información experimental y que han sufrido un avance importantísimo debido al uso de técnicas

automatizadas muy poderosas.

Por tanto, la cantidad de recursos disponibles para los investigadores es enorme, abarcando desde bases de datos de secuencias de genes y proteínas hasta ontologías biomédicas como *Gene Ontology (GO)* [Ashburner 2000] o bases de datos de artículos o literatura biológica y médica como *PubMed* [McEntyre 2001].

Es entonces cuando surge el problema ya comentado de la terminología, es decir, cómo relacionar los nombres de enfermedades, de genes, de proteínas, de funciones moleculares, etc, en los diferentes textos biomédicos con los diferentes vocabularios y nomenclaturas.

El Procesamiento de Lenguaje Natural (PLN) es una rama del campo de la Inteligencia Artificial que trata de ir más allá del reconocimiento léxico de palabras para interpretar texto comprendiendo la sintaxis (la gramática), la semántica (el significado) y más capas de análisis.

Con el PLN podemos dividir texto libre en párrafos, frases y componentes de las frases [Smith 2004] para luego encontrar conceptos al compararlos con los términos (expresión textual de un concepto). Luego, estos términos se expanden en sus variantes lingüísticas dando como resultando un conjunto enorme de palabras. Cada par frase-palabra se puntúa de acuerdo a conceptos como cohesividad, cobertura, centralidad y variación y como resultado final se presentan varios conceptos y su puntuación para cada una de las frases.

Incluso los más sofisticados programas de PNL no pueden hasta el momento competir con el tratamiento manual de los datos mediante el uso de curadores o revisores expertos, quizás ayudados, eso sí, por herramientas automáticas como la que se presenta en este trabajo.

El PLN muestra algunos problemas para la identificación de conceptos sobre textos libres. Primero si las palabras claves de un concepto se muestran en el texto en diferentes frases y dado que NLP suele trabajar primero detectando frases y sentencias para luego buscar los conceptos asociados a cada una de ellas [Zou 2003]. En segundo lugar los programas NLP suelen demandar gran cantidad de recursos de computación que no los hace idóneos para aplicaciones en tiempo real.

En este trabajo nos proponemos demostrar que utilizando técnicas alternativas al PLN [French 2009] y con un preprocesado de los datos de Ontologías [Nadkarni 2001] podemos soslayar los problemas comentados, sin renunciar ni a la certeza y ni a la eficiencia en las búsquedas, e incluso incrementado la precisión.

Para ello nos basaremos en el uso de los datos de la integración de Ontologías o Vocabularios Controlados del *Unified Modeling Medical System (UMLS)* [Bodenreider 2004], datos que preprocesaremos con algoritmos de "Stemming" para la eliminación de sufijos en palabras del lenguaje inglés [Porter 1980] y que guardaremos en tablas SQL (MySQL) indexadas que serán usadas en el proceso en línea de las preguntas en texto libre de los usuarios e investigadores ofreciendo en tiempo real los conceptos de la citada Ontología UMLS más relacionados con el texto objetivo.

Construimos pues una matriz cuyas filas son las palabras (sin sufijos) del texto objetivo y

las columnas cada uno de los conceptos con los que están relacionados. Para mostrar los conceptos ordenados al usuario, nos basamos en el concepto de Distribución Hipergeométrica para calcular un peso normalizado (“Weight” o “Score”) basado en las palabras que definen cada concepto y en su frecuencia de uso en los diferentes vocabularios, pudiendo ofrecerse conceptos parcialmente relacionados con el tema del texto presentado, al contrario que lo propuesto en el artículo sobre IndexFinder [Zou 2003] en el que se descartan los impactos incompletos; es decir que si en el nombre de un concepto de varias palabras, hay alguna palabra que no se encuentra entre las búsquedas inicialmente, ese concepto queda descartado, independientemente de si la palabra aporta poco o mucho del significado.

Dejamos también el camino abierto al tratamiento de las matrices dispersas (“*sparse*” o con muchos ceros) que genera el proceso FastUMLS actualmente con técnicas de *bi-clustering* u otras, para poner de manifiesto submatrices de términos y conceptos que muestren patrones de expresión comunes tal y como se hace en matrices de genes por palabras o genes por expresiones con SENT [Vazquez 2009].

Inicialmente la motivación del trabajo fue el procesar una lista de palabras relacionada con uno de los grupos de genes que agrupa la herramienta Bioinformática SENT. SENT utiliza la Factorización No Negativa de Matrices para identificar términos en artículos científicos relacionados con un conjunto de genes, y utiliza dichos términos para agrupar y relacionar dichos genes. SENT ofrece al investigador una forma de interpretar los datos de experimentos con gran cantidad de genes ofreciendo grupos y una lista de palabras que indiquen la relación entre ellos.

El tratar de inferir qué conceptos formales, de una base de datos revisada por expertos y con relaciones entre ellos, podrían estar relacionados con la lista de palabras clave ofrecida por SENT parecía un salto cualitativo en la información a mostrar al investigador.

Una vez tratada una lista genérica de palabras, el utilizar la misma casuística para procesar frases y textos biomédicos ha surgido como el siguiente paso natural que ha dado lugar a la herramienta FastUMLS y al presente trabajo, en el que se comparan resultados con herramientas de PLN y se muestran los datos de las revisiones de resultados por expertos. *InterActive MetaMap* es la herramienta de PLN con la que comparamos y que extrae conceptos de UMLS a partir de texto libre, y que utilizaremos para comparar resultados con los ofrecidos por la herramienta propósito de este trabajo.

Queda así introducida la función del proceso que se presenta para el procesamiento de texto genérico y la extracción o anotación automática de conceptos formales almacenados en el Tesoro de UMLS.

A continuación explicaremos algunos conceptos relacionados con las Ontologías y Tesoros en general y con UMLS en particular que nos ayudarán comprender los objetivos, los métodos y los resultados del trabajo realizado.

## Ontologías

Una Ontología aporta un mayor nivel de conocimiento sobre determinado dominio que un mero vocabulario o diccionario de términos dentro del área de conocimiento objetivo, y suele combinar varios esquemas en una estructura completa de datos que contiene todos los términos utilizados y las distintas relaciones entre ellos. Además una Ontología suele aportar atributos a cada término que nos permiten agrupar, indexar o ayudar en la búsqueda de información.

Según la Wikipedia, una Ontología o Tesauro es una lista que contiene los términos empleados para representar los conceptos temas o contenidos de los documentos de un dominio, con miras a efectuar una normalización terminológica que permita mejorar el canal de acceso y comunicación entre los usuarios.

Los términos de la Ontología se relacionan entre sí bajo tres modalidades de relación:

- Relaciones jerárquicas: Establece subdivisiones que generalmente reflejan estructuras de TODO/Parte.
- Relaciones de equivalencia: Controla la Sinonimia, Homonimia, Antonimia y Polinimia entre los términos.
- Relaciones asociativas: Mejoran las estrategias de recuperación y ayudan a reducir la polijerarquía entre los términos.

Aunque los incluye, las entradas de un tesauro no deben ser consideradas sólo como una lista de sinónimos.

Las Ontologías pueden tener una estructura extremadamente compleja o bien ser relativamente sencillas, pero lo más importante de ellas es que capturan conocimiento de un dominio de una forma que puede tratarse fácilmente con procesos informáticos [Stevens 2007]. Como los términos de una Ontología y las relaciones entre ellos están definidas de una forma clara – formalmente – el uso de ellas facilita tareas como el anotar adecuadamente los elementos y estandarizar nombres, mejorar los resultados de las búsquedas informáticas y servir de soporte a sistemas de inferencia de conocimiento a partir de los datos de la Ontología.

En la última década hemos visto un fuerte incremento en el número de sistemas para la representación de entidades de Biología Molecular y de Biomedicina, sus términos y sus relaciones. A menudo se han nombrado estos sistemas como Vocabularios, Terminologías, Tesauros (del inglés, a su vez del latín, “*Thesaurus*”) u Ontologías, y en la práctica puede decirse que se utilizan estos nombres de forma completamente intercambiable. En este trabajo utilizaremos indistintamente los citados términos.

El creciente interés en ellas se pone de manifiesto en el número de citas a Ontologías en los artículos de la base de datos más importante de publicaciones Biomédicas, PubMed/MEDLINE, que ha crecido en un seiscientos por cien por año [Bodenreider 2008] en los últimos años.

Una de las Ontologías más conocidas y usadas en Biología Molecular es *Gene Ontology* [GO]. *Gene Ontology* es un proyecto colaborativo que trata de resolver la necesidad de contar con descripciones consistentes de productos genéticos en diferentes bases de

datos. La comunidad de *Gene Ontology* desarrollan un mantienen tres Ontologías que describen estos productos de acuerdo con sus asociaciones con Procesos Biológicos, Componentes Celulares y Funciones Moleculares.

Pues bien, *Gene Ontology* se ha convertido en la Ontología más citada en los artículos sobre Biomedicina y Bioinformática con más de 450 citas por año [Bodenreider 2008] lo que pone de manifiesto el creciente interés y necesidad en este tipo de almacenes de información.

El esfuerzo por crear y mantener estas Ontologías es muy importante en tiempo y en recursos, y depende principalmente de los “curadores” (“curators” en inglés) o revisores, que son personas expertas en el dominio de conocimiento de la Ontologías encargadas de comprobar la corrección de los términos y de las relaciones entre ellos, muchas veces ayudados por herramientas automáticas – de la que la presentada en este trabajo puede ser una opción – pero aún así con todavía una ingente parte de trabajo rutinario de revisión de los artículos y literatura relacionada para la detección de nuevos términos o de conceptos redundantes, así como de la creación y mantenimiento de las relaciones entre los distintos elementos.

Algunas de las Ontologías más conocidas son:

- *SNOMED CT*, una guía de conceptos sobre medicina de las salud.
- El *Logical Observation Identifiers, Names and Codes (LOINC)* , un vocabulario de identificadores y nombres para ensayos clínicos.
- El *Foundational Model of Anatomy (FMA)*, una ontología sobre anatomía humana y las relaciones entre las diferentes partes del cuerpo humano.
- *Gene Ontology (GO)*, un vocabulario controlado para la anotación de productos genéticos cruzado entre especies y del que ya hemos hablado.
- *RxNorm*, un vocabulario con los nombres y códigos normalizados de medicamentos.
- El *National Cancer Institute Thesaurus (NCIT)* un diccionario de dominio público especializado en cáncer.
- El *International Classification of Diseases*, agrupación de terminología médica y de enfermedades con más de ciento quince años de antigüedad
- El *Medical Subjects Headings (MeSH)*, muy utilizado vocabulario para indexar, anotar y recuperar artículos y literatura científica en Biomedicina.
- El ***Unified Modeling Medical System (UMLS)*** [Bodenreider 2004] es un sistema unificado de terminología controlada que integra todas las Ontología mencionadas y muchas otras más, con el propósito de ofrecer conceptos únicos con referencias a cada vocabulario origen y relaciones entre ellos incluso entre términos cruzados de diferentes orígenes, que explicaremos con detalle más adelante y en el que se basa el anotador automático propuesto en este trabajo.

Otra Ontología de especial interés para la anotación de elementos biomédicos es la *Open*

*Biomedical Ontologies (OBO) [Smith 2007].*

Las Ontologías tienen diferentes aplicaciones, además de las ventajas obvias de organizar el conocimiento en un dominio determinado. Las Ontologías se utilizan para etiquetado formal de todo tipo de documentos o conjuntos de palabras, para mejorar las búsquedas, para encontrar nuevas relaciones entre elementos y para inferir conocimiento desconocido previamente cuando se usan en conjunto con Sistemas de Ayuda a la Toma de Decisiones [Greenes 2007] y también sirve de base a numerosos algoritmos de Procesamiento del Lenguaje Natural (PNL) con objetivos como la extracción de información, extracción de relaciones, resúmenes automáticos de documentos (objetivo parecido uno de los del programa FastUMLS que se presenta), respuestas automatizadas a preguntas y, en general, a todo lo relacionado con la minería de texto.

Prácticamente todas las Ontologías reseñadas se utilizan con el propósito de anotar o indexar textos científicos ya sean éstos artículos, descripciones de ensayos y experimentos, definiciones de proteínas, indicaciones de medicamentos o descripciones de síntomas de pacientes.

La indexación o anotación se refiere normalmente a la asignación de entradas de una Ontología a documentos o textos. De esta forma se enriquecen sustancialmente las bases de datos de documentos, quedando cada elemento anotado o etiquetado, pero de manera formal y normalizada, permitiendo que la búsquedas estén más orientadas al conocimiento que a la información en bruto.

Todavía hoy la anotación de las grandes bases de datos de documentos como PubMed/MEDLINE se sigue haciendo principalmente forma manual, y el campo de los anotadores automáticos es un área de investigación en auge por razones obvias.

Usando Ontologías con relaciones jerárquicas entre sus elementos tales como MeSH o UMLS, las búsquedas contra documentos indexados pueden expandirse usando los descendientes de los términos originales (pe. "lóbulo temporal" como parte de "cerebro") además de poder enriquecerse utilizando los sinónimos de dichos términos.

Las Ontologías y sobretodo algunas de ellas como UMLS, permiten acceder a mejores niveles de respuestas a preguntas que hasta ahora sólo se respondían utilizando las capacidades de búsqueda de texto como las que ofrece PubMed/MEDLINE o incluso Google-Scholar.

En las industrias biotecnológicas y las farmacéuticas hay cada vez más presiones para la identificación temprana de objetivos de medicamentos que ofrezca a su vez una ventaja competitiva en la arena de los negocios. No es por lo tanto sorprendente que en un estudio reciente se muestren datos del creciente uso de búsquedas en los textos científicos por parte de las empresas farmacéuticas y que el principal tema de las preguntas sean medicamentos, enfermedades, genes y proteínas [Agarwal 2008].

En este trabajo se utilizan los datos de UMLS [Bodenreider 2004] que integra a su vez otras importantes ontologías, para poder anotar y enriquecer un texto libre y abrir el camino a la indexación de grandes bases de datos, de la ayuda automática a revisores y para la inferencia de conocimiento basada en relaciones entre conceptos y en la búsqueda de grupos con significancia estadística.

## UMLS

El *Unified Medical Language System*® (UMLS®) [Bodenreider 2004] es un repositorio integrado de Vocabularios en los dominios de la Medicina y la Biología desarrollado por la *US National Library of Medicine* en un esfuerzo sostenido durante ya más de veinte años.

UMLS se desarrolló como una iniciativa para resolver dos problemas fundamentales de los procesos para la recuperación de información automática y basada en computadores: Primero la variedad de nombres utilizados referirse al mismo concepto y, segundo, la falta de un formato estándar para distribuir y difundir las terminologías.

Los principales objetivos de UMLS son el proporcionar un *middleware* para uso intelectual en entornos biomédicos, ofrecer un conjunto de herramientas para desarrolladores de sistemas y utilizar el conocimiento de diferentes vocabularios para resolver el problema de las disparidades en el uso del lenguaje (“contusión”, “hematoma”).

UMLS Integra más de 60 familias de Vocabularios médicos y biológicos, que abarcan desde enfermedades o terminología sobre anatomía hasta funciones moleculares de los genes y de las proteínas de diferentes especies. En su edición de 2009 la base de datos o Tesoro de UMLS alberga más de dos millones conceptos diferentes y más de 13 millones de relaciones entre ellos y contiene unos 3,2 millones de cadenas referenciadas con sus respectivos diccionarios o vocabularios orígenes.

De los tres componentes de UMLS el principal es el Tesoro o *Metathesaurus*®, repositorio de conceptos y sus relaciones.

Los otros dos elementos son:

- UMLS *Semantic Network*:

Proporciona un amplio espectro de categorías (153), o tipos semánticos, en el que se categorizan todos los conceptos del *Metathesaurus* y un conjunto de relaciones (54) que existen entre los citados tipos. Estas categorías nos sirven para clasificar los conceptos, buscar conceptos relacionados y establecer dependencias.

Cada concepto de UMLS tiene asociado al menos un tipo semántico de la jerarquía, y siempre se intenta asociar el tipo más específico al concepto. Por ejemplo, al concepto “Macaco” se le asocia el tipo semántico “Mamífero” porque no existe un tipo “Primate” más específico disponible en la jerarquía de tipos.

En el trabajo que se presenta no se ha hecho uso de la Red Semántica de UMLS, aunque sí acompañan los resultados de los conceptos relacionados con el texto objetivo con el tipo semántico principal, que puede usarse para filtrar o agrupar los resultados.

- *SPECIALIST Lexicon* y Herramientas Léxicas:

El *SPECIALIST Lexicon* ha sido desarrollado para proporcionar la información léxica que necesitan las herramientas de Procesamiento de Lenguaje Natural de *Lexicon* e incluye las palabras comunes del lenguaje inglés añadiendo el vocabulario biomédico.

Dentro de las herramientas que proporciona UMLS, una de las más usadas es

MetaMap y MetaMap Transfer Information o MMTx [Aronson 2001].

MMTx hace accesible el código y las herramientas de MetaMap para que puedan ser usadas por investigadores en el campo de la Bioinformática. Con MetaMap y mediante técnicas de PLN, podemos relacionar texto libre con conceptos de UMLS, objetivo que persigue también el proceso presentado en este trabajo pero como veremos, con diferencias importantes en cuanto a su especificidad, resultados y tiempo de procesamiento.

Las diferentes partes de UMLS pueden descargarse desde la Web o bien encargar el envío de un DVD previa aceptación de licencia de uso.

Hasta aquí la introducción que sirve para ubicar el trabajo en las áreas de la Bioinformática y del uso de Ontologías para la ayuda a resolver problemas biomédicos basados en la minería de texto.

Presentaremos en las páginas siguientes los conceptos técnicos en los que se basa FastUMLS para luego detallar el modelo de procesos y arquitectura software utilizada y terminar con una discusión de los resultados y la explicación de las líneas de trabajo futuras para las que este trabajo sirve de base.

## Conceptos y Terminología

El objetivo principal de FastUMLS es el tratamiento de una lista no ordenada de palabras, que puede ser un texto libre, para ofrecer un conjunto de conceptos normalizados que sirvan para comprender el contenido al que se hace referencia.

Estos elementos se recogen de la base de datos de UMLS en donde todo el contenido está organizado en torno al **concepto**, osea, por significado, y de cada concepto podemos mostrar información relacionada como:

- Cada una de las cadenas asociadas y sus agrupaciones léxicas (eliminando variaciones como por ejemplo las debidas a mayúsculas y minúsculas).
- En qué vocabularios o diccionarios origen podemos encontrarlas.
- Los tipos o categorías a las que pertenece un concepto.
- Qué conceptos están relacionados con uno determinado y de acuerdo a diferentes tipos de relación.

Recordando lo mencionado, en UMLS todos los términos que son sinónimos entre sí se agrupan en un único concepto gracias al trabajo de los revisores manuales, que cuando detectan un nuevo concepto en alguno de los vocabularios orígenes lo dan de alta, o lo eliminan o unifican si se detecta redundancia con un concepto previo.

Cada versión del *Metathesaurus* de UMLS se acompaña de un fichero (MRCUI) , o tabla SQL si se carga en una base de datos relacional, con los cambios de entradas y reducción de redundancias detectados que puede ser consultado para comprobaciones de cambios.

Es importante hacer notar que el *Metathesaurus* refleja y preserva los significados, nombres de los conceptos y las relaciones de los vocabularios origen.

Por ejemplo si el concepto “Dolor de Cabeza” aparece con diferentes denominaciones en dos diccionarios como “Dolor de Cabeza” en uno y “Cefalea” en otro, en el *Metathesaurus* tenemos un concepto asociado a dos entradas diferentes con diferente cadena y cada una relacionada con el diccionario adecuado. Si en un futuro se incorporara la definición “dolor craneal” en otro diccionario, aparecería una nueva entrada con la citada cadena asociado al mismo concepto que ya existía. De todas las cadenas o definiciones asociadas a un concepto o definición UMLS marca una como preferida.

Uno de los objetivos de UMLS es relacionar y conectar los diferentes nombres o formas de referirse a un concepto en los diferentes vocabularios. UMLS asigna diferentes tipos de identificadores únicos y permanentes a cada significado y cada nombre de concepto, además de preservar los identificadores que tengan en el vocabulario origen MeSH, GO o el que fuere. Toda la estructura de conceptos se ofrece en un fichero, o tabla SQL como es el caso de FastUMLS, llamado MRCONSO.

## **Conceptos e Identificadores de Concepto (CUIs)**

Como hemos dicho un concepto es un significado, y un significado puede tener diferentes nombres. Los “curadores” de UMLS se revisan los vocabularios y tratan de comprender el significado de cada nombre en cada vocabulario y enlazar todos los nombres de todos los vocabularios que significan lo misma cosa.

Cada concepto o significado en el *Metathesaurus* tiene un identificador único y permanente al que nos referiremos como CUI (“*Concept Unique Identifier*” o Identificador Único de Concepto).

## **Términos e Identificadores Léxicos (LUIs)**

Un término agrupa un conjunto de variaciones léxicas de la misma cadena. Por ejemplo, “adenoidectomy”, “ADENOIDECTOMY” y “Adenoidectomies” son diferentes cadenas que hacen referencia al mismo término y reciben en UMLS el identificador único léxico o LUI de L0001425 que a su vez hace referencia al concepto con identificador C0001425.

A su vez “Adenoidectomy without tonsillectomy” recibe un identificador de término o léxico (LUI) de L0001426 que hace referencia al mismo concepto y por lo tanto relacionado con el mismo CUI.

En este ejemplo, por un lado tenemos cuatro cadenas – o nombres diferentes en al menos una letra – que hacen referencia al mismo concepto, y dos términos que agrupan las diferentes variaciones léxicas.

## **Nombres de Conceptos e Identificadores de Cadenas (SUI)**

Cada nombre de concepto o cadena en cada lenguaje en el *Metathesaurus* tiene un identificador único y permanente o SUI.

Cualquier variación en las letras o conjunto de caracteres, de mayúsculas o minúsculas o de un signo de puntuación es una cadena diferente con diferente SUI. Si la misma cadena tiene más de un significado, el identificador SUI estará asociado a más de un concepto, como por ejemplo “cold” en inglés que significa “frío” o “resfriado” según el contexto.

## **Átomos e Identificadores de Átomos (AUI)**

Los átomos son el componente básico de la estructura del *Metathesaurus* y son los nombres de los conceptos o cadenas de cada uno de los vocabularios. Cada vez que una cadena aparece en un vocabulario se le asigna un identificador único o AUI (“*Atom Unique Identifier*”). Si la misma cadena aparece en muchos vocabularios diferentes se le asigna a cada ocurrencia un AUI distinto, y todos esos AUIs estarán relacionados con el mismo SUI.

En la Tabla 1 se muestran la única entrada relacionada con el concepto de CUI "C0009176" que hace referencia a la "Intoxicación por Cocaína", y que sólo aparece una sola vez en el diccionario "DXPlain" (DXP) del programa de diagnóstico experto del Hospital General de Massachusetts, por lo que sólo se relaciona un CUI, con un LUI, un SUI y un AUI.

STR	LUI	SUI	AUI	SAB
COCAINE INTOXICATION	L0009176	S0361244	A0397867	DXP

Tabla 1

Sin embargo en la Tabla 2 podemos comprobar que para el concepto de significado "Enfermedad de Addison" que es una deficiencia hormonal causada por daños a la glándula renal que ocasiona languidez y debilidad general, irritabilidad gástrica y cambio en la coloración de la piel, y que tienen como CUI "C0009176", aparecen 34 AUIs que son 34 cadenas diferentes cada una en un vocabulario (MeSH, NCI, MEDLINEPLUS, etc) y con sus correspondientes agrupaciones léxicas (10).

STR	LUI	SUI	AUI	SAB
Addison Disease	L0001403	S0010792	A6954527	MSH
Addison's Disease	L0001403	S0010794	A15661390	MEDLINEPLUS
Addison's Disease	L0001403	S0010794	A6954528	MSH
Addison's Disease	L0001403	S0010794	A7568543	NCI
Addisons Disease	L0001403	S0010796	A0019742	MSH
Disease, Addison	L0001403	S0033587	A0049628	MSH
ADDISON DISEASE	L0001403	S0352252	A0385543	DXP
ADDISON'S DISEASE	L0001403	S0352253	A0385544	CST
Addison's disease	L0001403	S0354372	A0388276	AOD
Addison's disease	L0001403	S0354372	A0388277	CSP
Addison's disease	L0001403	S0354372	A0388279	LCH
Addison's disease	L0001403	S0354372	A4367951	MTH
DISEASE ADDISON'S	L0001403	S0365923	A0404749	CST
Addison's disease NOS	L0001403	S1921523	A12975231	MTHICD9
ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)	L0278071	S0352321	A0385630	COSTAR
Adrenal insufficiency (Addison disease)	L0278071	S7557005	A11974577	OMIM
Chronic Adrenal Insufficiency	L0278357	S7669391	A12791935	NCI
ADRENOCORTICAL INSUFFICIENCY, PRIMARY FAILURE	L0278422	S0352329	A0385641	DXP
Primary Adrenal Insufficiency	L0494851	S5907334	A12807945	NCI
Primary Adrenal Insufficiency	L0494851	S5907334	A6975965	MSH

Adrenal Insufficiency, Primary	L0494851	S5924573	A6993206	MSH	
Insufficiencies, Primary Adrenocortical	L0494940	S5901878	A6970509	MSH	
Insufficiency, Primary Adrenocortical	L0494940	S5901881	A6970512	MSH	
Primary Adrenocortical Insufficiencies	L0494940	S5907335	A6975966	MSH	
Primary Adrenocortical Insufficiency	L0494940	S5907336	A6975967	MSH	
Adrenocortical Insufficiencies, Primary	L0494940	S5924576	A6993209	MSH	
Adrenocortical Insufficiency, Primary	L0494940	S5924577	A6993210	MSH	
Hypoadrenalism, Primary	L0585243	S5901432	A6970063	MSH	
Primary Hypoadrenalism	L0585243	S5907343	A6975974	MSH	
ADDISONS DIS	L6328093	S7256177	A12075312	MSH	
Hypocortisolism	L6572371	S7672732	A12791936	NCI	
Hypocortisolism	L6572371	S7672732	A15661692	MEDLINEPLUS	

Tabla 2

- Cada uno de los CUI (concepto) está enlazado al menos con un AUI (átomo), un SUI (cadena) y un LUI (agrupación léxica o término), pero puede estarlo con muchos.
- Un único AUI (átomo) está enlazado con un único SUI (cadena), un sólo LUI (término) y un sólo CUI (concepto).
- Un SUI (cadena) puede estar relacionado con muchos AUIs (átomos), sólo con un LUI (término) y con más de un CUI (concepto) aunque el caso típico será con sólo un concepto.
- Un LUI (término o agrupación de variaciones léxicas) puede estar relacionado con muchos AUIs (átomos), muchos SUI (cadenas) y normalmente con un sólo CUI.

En el ejemplo de la Figura 1, podemos ver la relación del concepto C0004238 con sus diferentes nombres (SUI), cómo éstos se agrupan en dos términos que eliminan las diferencias léxicas L0004238 y L0004327, y cómo las cadenas (SUI) se corresponden con uno o varios átomos (AUI) en los diferentes vocabularios.

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)	A0027665 Atrial Fibrillation (from MSH)  A0027667 Atrial Fibrillation (from PSY)
		S0016669 Atrial Fibrillations	A0027668 Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	S0016899 Auricular Fibrillation (preferred)	A0027930 Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	A0027932 Auricular Fibrillations (from MSH)

Figura 1 - Documentación Web UMLS

## Modelo de Datos UMLS

La información de UMLS se distribuye en un DVD que puede encargarse desde la Web o mediante descarga directa, si previamente se forma un acuerdo de licencia.

Los datos se distribuyen en ficheros de formato RRF y todos los nombres comienzan con las letras MR (*Metathesaurus Relational*) y siguen con caracteres que representan la información que contiene el fichero. Por ejemplo MRREL, contiene las relaciones entre conceptos, y MRSAB contiene el listado de vocabularios y sus abreviaturas, del inglés “*source abbreviations*”.

Estos ficheros pueden verse como tablas con un número fijo de columnas y un número variable de filas o registros que varía en cada distribución que suele hacerse una vez por año, y cada uno corresponde a uno de los cuatro grupos lógicos siguientes:

- Conceptos, Nombres de Conceptos y sus Orígenes. MRCONSO.
- Atributos. MRSAT, MRDEF, MRSTY y MRHIST
- Relaciones. MRREL, MRCOC y otros.
- Datos sobre el propio *Metathesaurus*. MRFILES, MRCOLS y otros.
- Índices y datos tratados para procesamiento de texto. MRXW\_ENG, MRXW\_SPA.

En la Tabla 3 se listan los principales archivos, junto con su descripción y el número de registros de la versión UMLS del año 2009.

Hemos resaltado los dos ficheros con los que trabaja el proceso FastUMLS para la extracción automática de conceptos desde texto libre.

Nombre de Fichero	Descripción en Inglés	Filas
MRCOC	Co-occurring Concepts	18.096.263
MRCOLS	Attribute Relation	308
<b>MRCONSO</b>	<b>Concept names and sources</b>	<b>3.276.579</b>
MRCUI	CUI History	1.014.829
MRDEF	Definitions	113.608
MRDOC	Typed key value metadata map	2.632
MRFILES	Relation Relation	45
MRHIER	Computable hierarchies	1.699.917
MRHIST	Source-asserted history	0
MRMAP	Mappings	260.972
MRRANK	Concept Name Ranking	297
MRREL	Related Concepts	13.513.394
MRSAB	Source Metadata	158
MRSAT	Simple Concept, Term and String Attributes	14.317.201
MRSMAP	Simple Mappings	31.113
MRSTY	Semantic Types	1.714.246
MRXNS_ENG	Normalized String Index	4.358.479
<b>MRXNW_ENG</b>	<b>Normalized Word Index</b>	<b>16.050.600</b>
MRXW_FRE	French Word Index	10.802
MRXW_GER	German Word Index	2.283
MRXW_HEB	Hebrew Word Index	1.617
MRXW_HUN	Hungarian Word Index	2.075
MRXW_ITA	Italian Word Index	10.291
MRXW_RUS	Russian Word Index	0
MRXW_SPA	Spanish Word Index	2.518
MRXW_SWE	Swedish Word Index	2.310

Tabla 3

## MRCONSO

En este fichero o tabla, hay exactamente una fila por cada átomo (AUI) – cada ocurrencia de un nombre de un concepto o cadena única en cada uno de los vocabularios origen – del *Metathesaurus*.

Las principales columnas de cada uno de los registros de MRCONSO son:

<b>AUI</b>	Identificador único de átomo (Cadena + Vocabulario)
<b>CUI</b>	Identificador único del concepto asociado al AUI.
<b>LUI</b>	Identificador único del término.
<b>SUI</b>	Identificador único de la cadena.
<b>SAB</b>	Nombre abreviado del vocabulario origen.
<b>CODE</b>	Identificador principal usado en el vocabulario origen. Muy útil para hacer referencia al origen.
<b>STR</b>	Cadena con el texto del AUI, SUI
<b>ISPREF</b>	Indica si es el átomo preferido para el concepto. La cadena que se supone que es más descriptiva y correcta.
<b>LAT</b>	Lenguaje de la cadena o término.
<b>TS</b>	Estado del Término

## MRXNW\_ENG

Esta tabla proviene del tratamiento de los datos de MRCONSO de forma que se procesa cada definición o cada cadena, normalizando palabras y dejando los términos en minúsculas para eliminar variaciones entre otros procesos, y sólo existe para el idioma inglés. NW proviene de palabras normalizadas o en inglés “*Normalized Words*”.

Primero se procesan los datos y se genera la tabla MRXNS\_ENG (NS = “*Normalized String*”) en la que se eliminan las palabras comunes o “*Stop Words*” en inglés, se normalizan las restantes y se ordenan las palabras dentro de una frase en orden alfabético. Esto permite que alguien procese de igual manera una definición y pueda comprobar directamente en esta tabla si hay una correspondencia directa.

En la tabla MRXNW\_ENG encontraremos una entrada para cada palabra normalizada en la cadena.

MRXNW\_ENG es tremendamente útil cuando queremos seleccionar qué conceptos, términos o cadenas de nombres contienen una determinada palabra como “cancer” o “Alzheimer”.

Veamos un ejemplo que ilustre el uso combinado de las tablas de conceptos, de cadenas normalizadas y de palabras normalizadas.

Por ejemplo para la cadena con SUI, S5907334 tenemos en MRCONSO:

STR	LUI	SUI	AUI	SAB
Primary Adrenal Insufficiency	L0494851	S5907334	A12807945	NCI
Primary Adrenal Insufficiency	L0494851	S5907334	A6975965	MSH

Si miramos en la tabla de cadenas normalizadas MRXNS\_ENG, encontramos las mismas palabras, en este caso en minúsculas y ordenadas alfabéticamente dentro de la cadena:

LAT	NSTR	CUI	LUI	SUI
ENG	adrenal insufficiency primary	C0001403	L0494851	S5907334

En la tabla con la palabras normalizadas y para este caso hayamos los siguientes registros:

LAT	NWD	CUI	LUI	SUI
ENG	adrenal	C0001403	L0494851	S5907334
ENG	insufficiency	C0001403	L0494851	S5907334
ENG	primary	C0001403	L0494851	S5907334

Si buscáramos conceptos en alguno de cuyos nombres o cadenas descriptivas aparezca la palabra “adrenal” iríamos a esta última tabla MRXNW\_ENG, preguntando si NWD es igual al “adrenal” y recuperaríamos un total de 2022 registros uno de los cuales sería el primero mostrado en la tabla anterior.

Como veremos en los métodos, FastUMLS hace uso de la tabla de Palabras Normalizadas durante la fase de preprocesado de los datos, en el que se tratan primero las palabras eliminando sufijos y variaciones que aportan una diversidad que no queremos, y posteriormente se precálculan pesos para los términos y para cada una de la palabras, ya sin sufijos, y así poder ponderar los resultados de los conceptos más susceptibles de interesar a un biólogo que quiere analizar un texto.

## MRSAB

En esta tabla tenemos un listado de todos los vocabularios incorporados en el *Metathesaurus* con información sobre cuándo se incorporaron al Tesauro, versión utilizada, contacto para tratar la información de uso y licencias, tipos de términos que encontramos o el lenguaje origen entre otros campos.

Algunos de los registros de la tabla pueden verse en la Tabla 4, junto con el número de términos que contiene MRCONSO de cada vocabulario.

RSAB	SON	LAT	TFR
AIR	AI/RHEUM, 1993	ENG	677
AOD	Alcohol and Other Drug Thesaurus, 2000	ENG	20685
AOT	Authorized Osteopathic Thesaurus, 2003	ENG	471
CCS	Clinical Classifications Software, 2005	ENG	1144
COSTAR	COSTAR, 1989-1995	ENG	3461
CSP	CRISP Thesaurus, 2006	ENG	21168
CST	COSTART, 1995	ENG	6410
DXP	DXplain, 1994	ENG	9974
FMA	Foundational Model of Anatomy Ontology, 2_0	ENG	133753
GO	Gene Ontology, 2008_04_01	ENG	82885
HCPCS	Healthcare Common Procedure Coding System, 2009	ENG	11504
HL7V2.5	HL7 Vocabulary Version 2.5, 2003_08_30	ENG	4973
HL7V3.0	HL7 Vocabulary Version 3.0, 2006_05	ENG	7678
HUGO	HUGO Gene Nomenclature, 2008_03	ENG	80036
ICD10PCS	ICD-10-PCS, 2008	ENG	299135
ICD9CM	ICD-9-CM, 2009	ENG	39215
ICPC	International Classification of Primary Care, 1993	ENG	1052
ICPCBAQ	ICPC, Basque Translation, 1993	BAQ	695
ICPCDAN	ICPC, Danish Translation, 1993	DAN	723
ICPCDUT	ICPC, Dutch Translation, 1993	DUT	723

Tabla 4 - Muestra de Vocabularios en UMLS

Cada uno de los vocabularios pertenece a una categoría en cuando al tipo de licencia de uso y es conveniente leer y comprender los términos de uso de cada uno en el acuerdo que se acepta al descargar u ordenar el DVD con los datos de UMLS.

El acuerdo de licencia puede encontrarse en el siguiente enlace:

<http://wwwcf.nlm.nih.gov/umlslicense/snomed/license.cfm>

La categoría 0 es la que no impone ninguna restricción al uso directo o derivado de los datos y existen cuatro categorías más cada cual más restrictiva.

## Metadatos sobre UMLS

Algunos datos sobre la versión de UMLS 2009AA, utilizada en este trabajo son:

Versión	:	2009AA
Formato	:	RRF
Conceptos	:	2.125.395
Nombres de Concepto (AUI)	:	9.691.753
Cadenas de Nombres de Concepto (SUI)	:	8.006.171
Términos o nombres normalizados (LUI)	:	7.246.004
Vocabularios	:	158
Palabras normalizadas en inglés	:	16.050.600
Idiomas contribuyendo con conceptos	:	19

## Métodos

### Objetivos

Como se ha explicado, FastUMLS de la motivación de tratar un conjunto de palabras clave provenientes de un tratamiento previo en la herramienta SENT, [Vazquez 2009]. Este conjunto de palabras no está ligado en un texto de una forma semánticamente coherente como ocurre en una descripción en lenguaje natural. Esto ha hecho que el acercamiento al tratamiento de textos libres por parte de FastUMLS haya sido diferente a la forma clásica de abordar este problema basándose en técnicas de Procesamiento de Lenguaje Natural, y ha permitido investigar métodos alternativos que son los que finalmente se han implementado.

Como la entrada a tratar no consistía en frases ordenadas y con un valor semántico y sintáctico, no se abordó el problema de la manera tradicional, sino que te enfrentó la situación de inferir qué conceptos de todos los posible podían tener mayor relación con las palabras de la entrada sin tener en cuenta su orden ni su relación en el texto o lista inicial.

Finalmente, FastUMLS se ha testado y probado contra textos en lenguaje natural demostrando precisión y eficacia en el tratamiento de los datos y en los resultados como veremos en la sección de resultados, y para ello se han añadido capacidades para el filtrado de palabras comunes o “*stop words*” y para la eliminación de sufijos o aplicación técnicas de “*stemming*”.

Establecido el objetivo principal, desde un primer momento se marcaron otros objetivos de carácter secundario atendiendo a la usabilidad de la herramienta y a su tipo de implementación:

- **Mostrar los datos de una forma amigable para el usuario:**

Muchas aplicaciones Bioinformáticas tienden a mostrar los datos de forma poco comprensible y usable por un usuario humano, a pesar de ser éste el usuario objetivo y el destinatario de los resultados. Presentar una lista lista de términos sobre un fichero XML de varios Megabytes de información es poco amigable para un usuario, sobretodo si tenemos en cuenta que el perfil en este caso es el de un Biólogo o un Médico, sin formación previsible en temas informáticos.

La salida de FastUMLS se presenta de forma legible en una tabla ordenada de conceptos, de forma que un usuario pueda identificar claramente los conceptos y centrarse, por ejemplo, sólo en los diez primeros. Este interfaz se sigue y se seguirá mejorando en siguientes pasos de este trabajo.

- **Capacidad para Interaccionar con otras aplicaciones:**

Además del interfaz de usuario, FastUMLS está preparado para enviar en un fichero XML todos los conceptos encontrados, que puede ser del orden de cadenas de miles, junto con su puntuación y posición relativa utilizando Web Services.

También puede utilizarse como proceso dentro de un sistema operativo para procesamiento Batch de textos de entrada.

- **Rendimiento: Utilizable para procesamiento On-line o en tiempo real.**

El modelo de procesos y de datos se ha estudiado para poder devolver resultados en tiempos humanamente razonables. Para ellos se han optimizado procesos, sentencias de bases de datos y tamaños de los datos utilizados.

- **Libre de ataduras respecto a sistemas operativos, plataformas y tecnologías:**

Con el interfaz Web independizamos la capa de uso desde la red para usuarios y con los Web Services, establecemos un interfaz claro para procesos, cualquiera que sea su tecnología.

Para la parte de procesamiento de los conceptos, hemos utilizado un Framework de desarrollo que permite desarrollar una vez y compilar para las principales plataformas de sistemas operativos y arquitectura de ordenadores como son Linux (en sus diferentes sabores), Windows, Solaris, HP-XU y AIX, y el lenguaje principal utilizado ha sido C++, atendiendo también a los objetivos de rendimiento.

Como motor de base de datos relacional se ha utilizado software de código abierto como es MySQL que además corre en las plataformas mencionadas.

Respecto a los conceptos ofrecidos la idea principal en la que se basa todo el trabajo es la siguiente, sabiendo que parto de una lista de palabras y que cada una está relacionada con un número más o menos grande de los términos de UMLS, es la siguiente:

¿Cuál es la probabilidad de que hubiéramos seleccionado cualquiera de los conceptos ofrecidos como resultado atendiendo al número de veces que cada palabra que pertenece al nombre del concepto está relacionada con otros conceptos, y teniendo en cuenta cuántas palabras de la entrada están también en la definición del concepto, y si esas palabras son muy comunes o no?

Dicho de otro modo, lo que queremos tener en cuenta la promiscuidad de las palabras a la hora de estar relacionadas con las cadenas que dan significado a los términos. Un concepto que tiene asociadas palabras muy genéricas – palabras que están presentes en muchos otros conceptos – tiene mayor probabilidad de salir que otro relacionado con palabras menos generales.

Basándonos en esta idea se precálculan los pesos relativos de los términos atendiendo a la número de veces que aparecen sus palabras en la base de datos, y en cada procesamiento se calcula el peso final dependiendo de qué palabras han impactado en qué conceptos.

De esta manera mostramos primero los conceptos asociados a términos que tienen mayor significancia estadística primando la especificidad frente a la generalidad de las definiciones.

## Arquitectura de Procesos

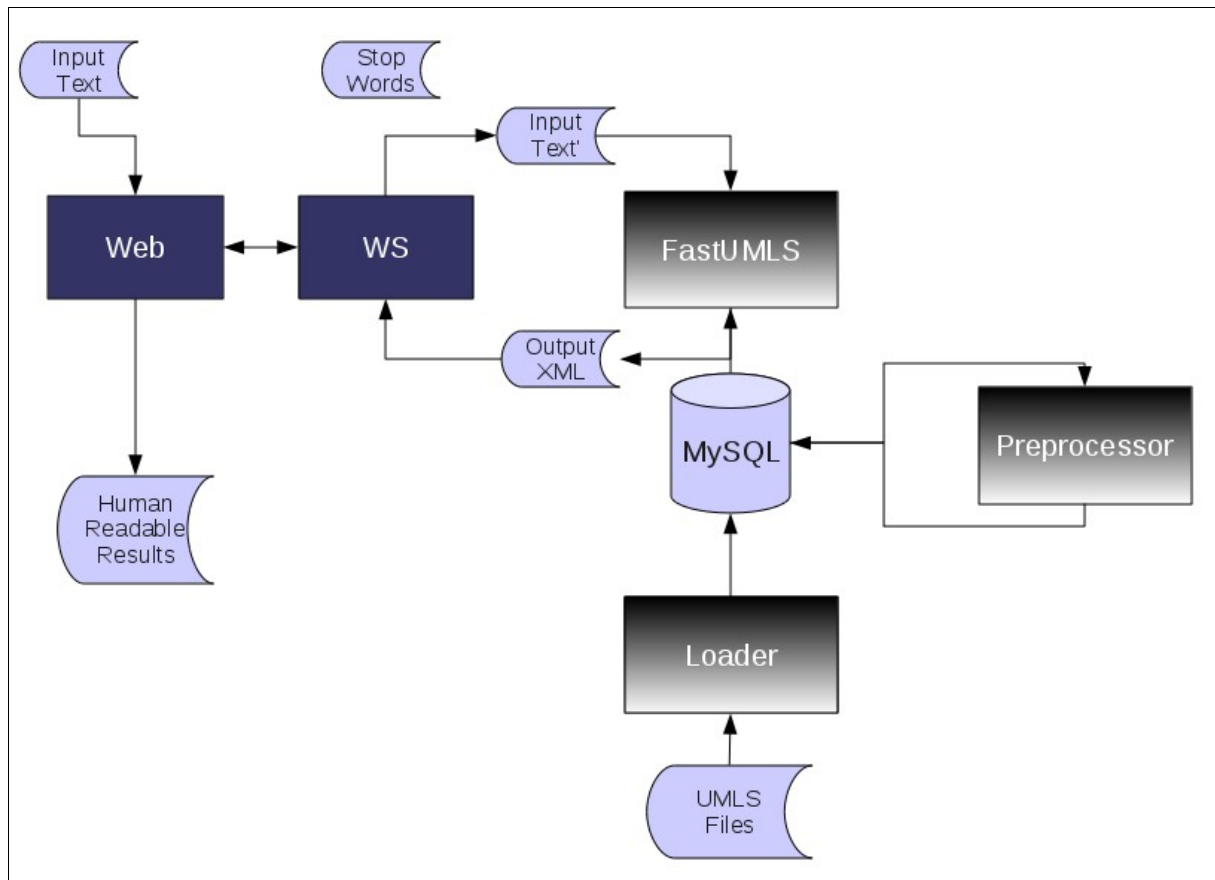


Figura 2 - Arquitectura de Procesos

Empezando desde la parte de abajo de la Figura 2 repasemos los diferentes procesos que componen la herramienta FastUMLS.

- **Carga Inicial o Loader:**

Como hemos comentado los datos del Metathesaurus se encuentran en formato RRF y necesitamos cargarlos en una base de datos, para facilitar el tratamiento en integración con otras tecnologías. Los datos vienen preparados para poderse cargar en tablas relacionales de varios motores de bases de datos uno de los cuales es MySQL y el utilizado en este caso.

Este proceso sólo se ejecuta una vez y su salida son los diferentes tablas e índices con los datos del Tesauro.

- **Preprocesador:**

Mediante este proceso que sólo se ejecuta una vez se transforman los datos de UMLS en otros preparados para que podamos procesar los datos adecuadamente según la metodología elegida.

En concreto, se transforma la tabla MRXNW\_ENG, con las palabras asociadas a cada término (LUI) y concepto (LUI) en otra en que hemos agrupado raíces de

palabras eliminando los sufijos (stemming) [Porter 1980] y hemos calculado cardinalidades de las palabras para precalcular el peso de los conceptos.

- **FastUMLS:**

Este es el proceso por lotes encargado de procesar un fichero con la lista de palabras a tratar, eliminar los sufijos de igual forma que el preprocesado, conectarse a la base de datos y devolver un fichero XML con los conceptos ordenados por peso según la cantidad de palabras concordantes con cada LUI (término) y su relativa promiscuidad.

Puede invocarse independientemente de los procesos Web.

- **Web Service:**

Este proceso permite utilizando interfaces basados en Web Services, el envío de una lista de palabras o un texto, y la recepción del fichero XML con los resultados.

Antes de llamar al programa por lotes de FastUMLS, elimina las palabras comunes o “stop words”, trata la aparición de caracteres como guiones o paréntesis, y formatea la entrada como una lista de palabras.

- **Interfaz Web:**

Permite que el usuario introduzca el texto a tratar e inicia la cadena de procesos llamando el Web Service.

(\*) Tanto el [Interfaz Web](#) como el [Web Service](#) han sido desarrollados por Rubén Nogales del Departamento de la Unidad de Bioinformática y Biología Computacional del CNB y la Universidad Complutense de Madrid.

## Preprocesado de los Datos

Ya hemos comentado que una de las tablas más importante para encontrar conceptos (CUI) o términos (LUI) asociados con una palabra, dentro de UMLS es **MRXNW\_ENG** en la que encontraremos una entrada para cada palabra normalizada en la cadena del nombre del término.

Utilizamos mejor el término o LUI en lugar del concepto, porque los que no nos interesa mezclar definiciones, como puede ocurrir con un concepto. Por ejemplo el concepto “dolor de cabeza” puede tener tres términos diferentes “Migraña”, “Dolor Cabeza”, y “Dolor Craneal”. Lo que queremos es tratar cada cadena separadamente buscando correspondencias de palabras en cada una, y posteriormente mostrar el concepto asociado.

Además y como se ha explicado, necesitamos procesar los datos para otorgarles un peso a las palabras que valore su promiscuidad, y otro a los términos basándonos en las promiscuidades que forman parte de sus cadenas o nombres. Como veremos estos pesos se basarán en las frecuencias de aparición no de las palabras sino de las raíces procesadas para eliminar sufijos mediante técnicas de “*stemming*” basadas en el algoritmo de Porter al que ya se ha hecho referencia.

Así pues tenemos como entrada de éste proceso la tabla **MRXNW\_ENG** cuyos campos son:

<b>LAT</b>	Idioma del término que en este caso siempre es el inglés. Podemos obviar este campo.
<b>NWD</b>	Palabra normalizada de la cadena asociada al término y al concepto. Ya se han eliminado plurales y algunas palabras frecuentes.
<b>CUI</b>	Identificador único de concepto.
<b>LUI</b>	Identificador único del término.
<b>SUI</b>	Identificador único de la cadena.

La idea con la tabla **MRXNW\_ENG** es la de hacer la siguiente query de SQL:

```
SELECT LUI
FROM MRXNW_ENG
WHERE NWD = 'cancer';
```

Y de esta forma recuperaremos todos los LUI de términos asociados a la palabra “cáncer”, para después recuperar la información que necesitamos de la tabla **MRCONSO**, por ejemplo utilizando esta otra sentencia:

```
SELECT CUI,SAB, CODE, STR
FROM MRCONSO
WHERE LUI = 'L8600667';
```

Obteniendo este resultado (en este caso un sólo registro):

```
+-----+-----+-----+-----+
| CUI      | SAB | CODE      | STR                                     |
+-----+-----+-----+-----+
| C2605685 | MSH | C533103   | highly express in cancer 1 protein, mouse |
+-----+-----+-----+-----+
```

Pero lo que queremos con nuestro proceso es primero eliminar la dispersión por sufijos de las palabras del campo NWD, y además calcular pesos y ordinales para palabras y conceptos, en forma de la tabla de nombre **FU\_MRXSNWCONSO\_ENG** y con campos:

<b>LUI</b>	Identificador único del término.
<b>SNWD</b>	Palabra normalizada y con Stemming
<b>LUI_WEIGHT</b>	Peso del LUI
<b>SNWD_WEIGHT</b>	Peso de la palabra

Así pues el objetivo del Preprocesamiento es el de generar como salida la tabla **FU\_MRXSNWCONSO\_ENG** a partir de las dos tablas originales MRCONSO y MEXNW\_ENG.

Una vez obtenida la tabla, en el proceso batch de FastUMLS hay que ir preguntando por los LUI relacionados con cada palabra junto con el peso del término por un lado y de la palabra por otro, e ir guardando cada uno y calculando el peso por impacto, para luego mostrar la información ordenada en memoria y generar el fichero XML como veremos.

Para llegar a obtener la tabla objetivo hemos tardado unas 18 horas de proceso, en el entorno de producción, siguiendo estos pasos:

- **Stemming o procesado de Sufijos:**

Utilizando el algoritmo de Porter se procesan todas las palabras dejando sus raíces. Para ellos el proceso batch de FastUMLS se ha preparado con un parámetro para que lo haga en una tabla auxiliar utilizando el mismo algoritmo que se usará cuando se procesen la palabras de la lista.

```
fastumls $ bin/fu_term2concept -c cfg/FASTUMLS.ini -S -d
```

- **Cálculo del peso de cada palabra sin sufijos:**

Para ello primero se calculan y se guardan cuántos términos diferentes están relacionados con cada palabra. Posteriormente se calcula el peso de la palabra (SNWD) de acuerdo a la siguiente fórmula logarítmica:

$$w_p = |\log_{10}(P_p)| \text{ (Campo SNWD\_WEIGHT)}$$

Siendo  $P_p$  la probabilidad de sacar la palabra “p” por azar del conjunto de filas de la tabla, es decir su cardinalidad o número de veces que aparece en la tabla asociada a un LUI, dividido por el total de registros de la tabla.

- **Cálculo del peso de cada término:**

Se calcula el peso normalizado de cada término atendiendo al peso de cada una de las palabras que le dan nombre (En “apoptosis celular”, usaríamos los pesos de ambas palabras para calcular el del término):

Sería la raíz cuadrada de la suma de los cuadrados de los pesos de sus palabras:

$$w_{LUI} = \text{SQRT} (w_{p1}^2 + w_{p2}^2 + \dots + w_{pn}^2 )$$

## Procesamiento On-line

Una vez generada la tabla **FU\_MRXSNWCONO\_ENG** podemos pasar al proceso de generación de conceptos, que se basa en la lista de palabras que introduce el usuario, u preprocesada por el Servicio Web, que elimina palabras comunes y algunos caracteres.

- **Tratamiento inicial**

El usuario introduce desde la Web un texto genérico, que es procesado para eliminar las palabras comunes del inglés de una lista de “stop words”, y además se eliminan caracteres que no se utilizan en las tablas indexadas de UMLS (MRXNW\_ENG) como los paréntesis, los guiones o las comillas.

- **Lanzamiento del Proceso de generación de conceptos:**

Se genera entonces un fichero con la lista de palabras a tratar, se asigna un número al proceso que se va a lanzar y un nombre para el fichero XML de salida, y se le pasa el fichero de configuración con la cadena de conexión a la base de datos.

El Servicio Web, queda en espera a que termine el proceso para recoger el fichero con la salida y posteriormente, éste se muestra al usuario.

La llamada al proceso sin parámetros puede verse en la Tabla 5.

```
FastUMLS 1.0
fu_term2concept -c config_file -i input_file -o outputfile
                 [-t min_terms] [-w min_weight] [-d] [-h]
... or ...
fu_term2concept -c config_file -S [-d] [-h] To Stem field SNWD

-c path to config file
-i path to input file. One term, one line
-o path to output file.
-t Minimum number of hits terms to print a concept. Optional, default 2.
-d Debug mode on. Optional
-h Print this info. Optional

FastUMLS is Open Source Code under GPL License
FastUMLS is based on ACE C++ frameworks and uses MySql client C API
ACE version is ACE 5.6.7
Author: Jose Luis Marina
http://jlmarina.net
BioInformatics group at Universidad Complutense Madrid - Spain
http://bioinfo.dacya.ucm.es
```

Tabla 5 - Texto de ayuda del proceso de de generación de conceptos

- **Proceso de generación de conceptos:**

El proceso de generación de conceptos sigue a su vez los siguientes pasos:

- ✓ Preprocesa cada palabra del fichero de palabras de entrada, aplicando el algoritmo de Porter para Stemming de palabras inglesas. Puede verse el autor e historial del código en la Tabla 6.
- ✓ Crea un Array en memoria con cada una de las palabras únicas sin sufijos

controlando repeticiones.

- ✓ Por cada palabra del Array anterior:
  - Recuperamos todos los conceptos asociados hasta un máximo.
  - Si el concepto es nuevo se añade a un Array de Conceptos, y si está repetido se tiene en cuenta el peso de la palabra para siempre ir calculando el peso de cada concepto en este segundo Array.
- ✓ Se ordena el Array de Conceptos en memoria y de acuerdo a su peso.
- ✓ Se genera el fichero XML trayendo además información sobre cadenas, vocabularios origen y código o forma de referencias en el origen.

```

/*****          stem.c          *****/

Purpose:   Implementation of the Porter stemming algorithm documented
           in: Porter, M.F., "An Algorithm For Suffix Stripping,"
           Program 14 (3), July 1980, pp. 130-137.

Provenance: Written by B. Frakes and C. Cox, 1986.
            Changed by C. Fox, 1990.
            - made measure function a DFA
            - restructured structs
            - renamed functions and variables
            - restricted function and variable scopes
            Changed by C. Fox, July, 1991.
            - added ANSI C declarations
            - branch tested to 90% coverage
            Changed by Arjen P. de Vries, March, 1996
            - fixed mismatch between prototype and function
            - removed unnecessary includes

Notes:     This code will make little sense without the the Porter
           article. The stemming function converts its input to
           lower case.

**/

```

Tabla 6 - Cabecera del fichero de código utilizado para el algoritmo de Porter

En el fichero de salida se muestran primero el array de términos de la siguiente manera:

```

<termArray>
<num_elements>5</num_elements>
  <term>
    <index>0</index>
    <name>gen</name>
    <num_concepts>372</num_concepts>
    <tot_concepts>369</tot_concepts>
  </term>
  <term>

```

```

    <index>1</index>
    <name>apoptosi</name>
    <num_concepts>602</num_concepts>
    <tot_concepts>574</tot_concepts>
  </term>
  <term>
    <index>2</index>
    <name>cell</name>
    <num_concepts>1200</num_concepts>
    <tot_concepts>24786</tot_concepts>
  </term>
  <term>
    <index>3</index>
    <name>cancer</name>
    <num_concepts>1200</num_concepts>
    <tot_concepts>5943</tot_concepts>
  </term>
  <term>
    <index>4</index>
    <name>xxnxxnxxn</name>
    <num_concepts>0</num_concepts>
    <tot_concepts>0</tot_concepts>
  </term>
</termArray>

```

A continuación en el fichero, encontraremos el Array de conceptos que para el caso de las cuatro palabras mostradas en el caso del Array de términos anterior contiene 3360 elementos y para cada concepto se recupera y se almacena la siguiente información:

- **Posición en el Array** : pe. 0
- **CUI y LUI del término** : pe. C1704351:L1222637
- Enlace a Web con información sobre el CUI.
- **Peso del concepto** : pe. 4.585077
- **Peso normalizado** : No se usa.
- **Número de palabras en la definición** : pe. 10
- **Número de palabras encontradas** : pe. 3
- Para cada una de las cadenas de nombre relacionadas con el concepto:
  - **Cadena** : pe. "Generator Dosing Unit"
  - **Identificador de Átomo o AUI** : pe. A10806099
  - **Código del Vocabulario Origen** : pe. NCI

- Código en origen : pe. C48496
- Para cada una de la palabras que trajeron el concepto:
  - Palabra : pe. Gen
- Un mapa bits con la palabras que trajeron el concepto:
  - Mapa de Bits : pe. 1 0 0 0 0

Puede verse un ejemplo del formato de salida en la Tabla 7.

```
<concept>
  <index>3</index>
  <cui>C0162638:L0189749</cui>
  <norm_weight>5.703012e-01</norm_weight>
  <weight>4.393189</weight>
  <num_terms>1</num_terms>
  <tot_terms>1</tot_terms>
  <strings>
    <string>
      <content>Apoptosis</content>
      <aii>A10763605</aii>
      <sab>NCI</sab>
      <code>C38784</code>
    </string>
    <string>
      <content>apoptosis</content>
      <aii>A11607855</aii>
      <sab>G0</sab>
      <code>G0:0006915</code>
    </string>
  </strings>
  <concept_terms>
    <term>apoptosi</term>
  </concept_terms>
  <terms_bitmap>
    0 1 0 0 0
  </terms_bitmap>
</concept>
```

Tabla 7 - Ejemplo de salida para concepto de FastUMLS

Por limitaciones de memoria no se cargan para cada término más de un número de conceptos que puede estar entre 1.000 y 5.000 en función de un parámetro de compilación.

Esto puede dejar fuera de la lista de términos algunos conceptos sin incorporar, pero no hemos detectado que supusiera un problema en los resultados, ya que palabras que estén relacionadas con más de 1.000 ó 2.000 conceptos, son consideradas por los algoritmos de cálculo de peso como bastante promiscuas y no suelen incidir en los resultados generales del proceso.

## Resultados y Conclusiones

Analizamos a continuación cuáles han sido los resultados de experimentos basados en FastUMLS, primero desde el punto de vista de capacidad de proceso y posteriormente se expondrán los resultados de cara a los objetivos y a la exactitud de los conceptos anotados.

Para comprobar si FastUMLS resulta ser una herramienta útil para la anotación de textos se han realizado dos tipos de prueba formal (además de las más informales buscando la valoración por impresiones de biólogos o expertos en textos biomédicos en las primeras fases del desarrollo).

Primero se han comparado resultados con la herramienta “*Interactive MetaMap*” (IM) [I-METAMAP] que utiliza las mismas bases de datos del MetaThesaurus y MetaMap para extracción de conceptos basándose en técnicas de PLN, procesando los mismos textos y comparando los conceptos recuperados por una y otra herramienta para el texto completo primero y posteriormente para cada una de las frases.

En segundo lugar se pidió la colaboración de cinco investigadores en Biomedicina y Biología Molecular, para que analizaran los resultados de FastUMLS después de haber procesado los resúmenes de artículos que previamente cada investigador había elegido, y de los que mayoritariamente era autor o colaborador, y por lo tanto experto en el material analizado.

### **Tiempos y Capacidad de Cálculo**

FastUMLS es capaz de procesar un texto de 200 palabras, contrastarlas contra la base de datos de 14 millones de registros, para recuperar conceptos y cadenas asociadas de entre más de 3 millones de elementos en tiempos de menos de 20 segundos corriendo sobre un servidor medio con 4 CPUs Intel Xeon a 3,40 GHz. Si queremos tratar la docena de palabras de una frase típica los tiempos se reducen a uno o dos segundos con el procesado de unos 5.000 conceptos.

Como hemos explicado, el FastUMLS genera una matriz de hasta 1200 conceptos por palabra, y en el caso de un texto del tamaño citado se tratan y ponderan unos 50.000 conceptos en total. Esta información se genera en un fichero XML con los conceptos ordenados según su peso estadístico y que puede ocupar 60 MBytes.

En el caso del proceso Web orientado a un usuario humano, se muestran los primeros conceptos hasta un total de 100, pudiendo variarse este número de forma parametrizada. Si se quisiera procesar la matriz completa para aplicar técnicas de *clustering* u otro tratamiento matricial, FastUMLS puede utilizarse a través del “*Web Service*” publicado o bien como proceso “batch” o por lotes en cualquier computadora en la que se haya compilado el proceso principal.

Los tiempos de procesamiento de FastUMLS han resultado al menos adecuados para uso en análisis de textos y frases en tiempo real, que era uno de los objetivos buscados, en contraposición a la mayoría de los programas de PLN que demandan mayores capacidades en cuanto a potencia de cálculo y suelen entregar los resultados de forma diferida en el tiempo.

## Pruebas Comparativas

La herramienta *Interactive MetaMap* (IM) la proporciona la *National Library of Medicine* del *National Institute of Health* (NIH) junto con los datos de UMLS para ayudar en la detección de conceptos del *Metathesaurus* en textos científicos Biomédicos.

El objetivo de estas pruebas es el de comprobar si la herramienta desarrollada, FastUMLS, proporciona un conjunto de conceptos comparables a una herramienta existente y utilizada por investigadores científicos [Aronson 2001] en el campo de la minería de textos Médicos y de la Biología. A pesar de que IM presenta determinados fallos como se muestra en el artículo de Divita, "Análisis de Fallos en MetaMap Transfer" [Divita 2004] y que se mantiene en constante evolución, es una referencia a la hora extraer términos de texto científico libre y es citado y utilizado en numerosos trabajos de investigación.

Para realizar las pruebas comparativas hemos utilizado anotaciones hechas por expertos sobre algunos genes de Levadura. Estos genes, su descripción y la literatura científica asociada ha sido revisada por expertos para finalmente asociar un conjunto de conceptos de *Gene Ontology* [GO].

El análisis de los textos asociados a uno de los genes ha servido para realizar dos comprobaciones:

- ¿Es capaz FastUMLS de extraer alguno de los conceptos anotados por humanos expertos en la materia?
- ¿Son comparables los resultados ofrecidos por FastUMLS e IM?

Para responder a la segunda pregunta ha habido que adaptar la entrada al tipo de procesamiento que realiza IM. *Interactive MetaMap* trata los textos desde el punto de vista del Procesamiento de Lenguaje Natural, y en este sentido, hace un trabajo previo en el que identifica párrafos dentro del texto, y dentro de cada uno selecciona frases y grupos semánticos utilizando analizadores léxicos y semánticos.

Es sólo una vez identificadas dichas frases cuando IM aplica la identificación de conceptos. Al contrario que IM, FastUMLS procede a la identificación de conceptos en todo el texto utilizado como entrada, con algunos inconvenientes (Ver Mejoras en el apartado de Líneas Futuras) , pero con bastantes ventajas como veremos.

Por lo tanto para comparar resultados entre las dos herramientas se ha tratado primero el mismo texto de descripción de un Gen, para luego ir procesando con FastUMLS cada una de las frases identificadas por IM, y, ahora sí, comparar los conceptos identificados.

## Identificación de Conceptos en Genes Anotados por Expertos

Yeast Genome o SDG (*Saccharomyces Genome Database*) es una colección organizada de información de biología molecular y genética sobre *Saccharomyces cerevisiae*, la levadura de la cerveza muy utilizada en investigación.

SDG contiene las secuencias de los genes y proteínas, descripciones y clasificaciones del papel biológico de cada una, sus funciones moleculares y localizaciones sub-celulares, además de enlaces a literatura, enlaces a conjuntos de datos sobre genómica y

herramientas para análisis y comparación de secuencias.

Dentro de SDG encontramos “Locus Page” con toda la información relacionada con cada Gen de la levadura, con el nombre y los sinónimos del Gen, anotaciones manuales sobre *Gene Ontology [GO]*, y descripciones de los genes y sus productos, así como una colección de artículos revisados o “curados” por expertos en la materia.

**Prueba 1: Gen RAP1**

Descripción del GEN:

*RAP1* (Repressor Activator Protein) encodes an essential protein involved in many diverse, some seemingly contradictory, processes in *S. cerevisiae*, including [telomere maintenance](#), transcriptional [silencing](#) (repression) of the silent mating loci *HML* and *HMR*, and high level [transcriptional activation](#) of genes encoding ribosomal proteins and glycolytic enzymes. In these various roles, the underlying function of *Rap1p* is to [bind DNA](#) in a sequence specific manner, often regulating the chromatin structure in the region where it binds (7, 8). *Rap1p* binds extensively in telomeric regions, where its function is related to both transcriptional silencing and telomere maintenance. In its role as a transcription activating factor, the largest group of target genes are those that encode ribosomal proteins. In rapidly growing yeast cells, the transcription rate of these genes is extremely high, accounting for about twenty percent of the total mRNA content of the cell. *Rap1p* is known to be required for the transcription of several non-ribosomal protein genes, including *HIS4* , *ENO1* and *ENO2* (12), and is implicated in transcriptional regulation of 185 additional genes.

Está anotado con los siguientes conceptos en Gene Ontology:

- [chromatin silencing \(IMP\)](#)
- [chromatin silencing at telomere \(IMP\)](#)
- [protection from non-homologous end joining at telomere \(IMP\)](#)
- [telomere maintenance via telomerase \(IMP\)](#)
- [transcription from RNA polymerase II promoter \(IMP\)](#)

Procesando todo el texto por ambas herramientas y buscando en IM en todo el texto que ofrece como resultado y en FastUMLS mirando sólo en los veinte primeros conceptos ofrecidos:

<b>FastUMLS</b>	<b>2 conceptos identificados</b> Dentro de los 20 primeros resultados. GO:0006342: chromatin silencing GO:0006348: chromatin silencing at telomere
<b>Interactive MetaMap</b>	<b>0 conceptos identificados.</b>

Respecto a los tiempos de procesamiento y aunque ambas herramientas resultan aceptables para el procesamiento en línea de los datos, FastUMLS es varias veces más rápido en mostrar los resultados.

Terms related	Weight
(GO) <a href="#">GO:0030466</a> : chromatin silencing at HML and HMR (sensu Saccharomyces)	8.774443
(GO) <a href="#">GO:0030466</a> : chromatin silencing at silent mating-type cassette (GO) <a href="#">GO:0030466</a> : chromatin silencing at silent mating-type cassette (sensu Fungi)	8.254469
(GO) <a href="#">GO:0006344</a> : maintenance of chromatin silencing	8.121110
(GO) <a href="#">GO:0006348</a> : chromatin silencing at telomere	7.604862
(GO) <a href="#">GO:0030466</a> : chromatin silencing at silent mating-type cassette (GO) <a href="#">GO:0030466</a> : chromatin silencing at silent mating-type cassette (sensu Fungi)	7.399025
(GO) <a href="#">GO:0031940</a> : activation of chromatin silencing at telomere	7.311663
(GO) <a href="#">GO:0031938</a> : regulation of chromatin silencing at telomere	7.202315
(GO) <a href="#">GO:0031940</a> : stimulation of chromatin silencing at telomere	7.056285
(GO) <a href="#">GO:0031939</a> : inhibition of chromatin silencing at telomere	6.994311
(GO) <a href="#">GO:0031940</a> : up regulation of chromatin silencing at telomere (GO) <a href="#">GO:0031940</a> : up-regulation of chromatin silencing at telomere (GO) <a href="#">GO:0031940</a> : upregulation of chromatin silencing at telomere	6.939218
(GO) <a href="#">GO:0031939</a> : down regulation of chromatin silencing at telomere (GO) <a href="#">GO:0031939</a> : down-regulation of chromatin silencing at telomere (GO) <a href="#">GO:0031939</a> : downregulation of chromatin silencing at telomere	6.937510
(GO) <a href="#">GO:0043007</a> : maintenance of rDNA (GO) <a href="#">GO:0043007</a> : rDNA maintenance (GO) <a href="#">GO:0043007</a> : ribosomal DNA maintenance	6.816663
(GO) <a href="#">GO:0006342</a> : chromatin silencing (MTH) NOCODE: chromatin silencing (GO) <a href="#">GO:0006342</a> : TCS	6.802456

Muestra de Resultados de FastUMLS para Gen RAP1

Es importante reseñar que FastUMLS busca conceptos no sólo del vocabulario GO, que es uno más de los incluidos en el *Metathesaurus* de UMLS.

## **Prueba 2: Gen CLN2**

Descripción del Gen:

CLN2 encodes a G1 cyclin involved in regulation of the cell cycle. Progression through the cell cycle is a carefully regulated process that is conserved throughout eukaryotes. Periodic activation of cyclin dependent kinases are required for this process; the critical CDK involved in cell cycle progression in yeast is Cdc28p. Cyclins are the regulatory subunits that activate CDKs at the appropriate time in the cell cycle; they were named for their cyclical accumulation during particular phases of the cell cycle. Distinct CDK-cyclin complexes are required for progression through different stages of the cell cycle. CLN1, CLN2, and CLN3 encode the yeast cyclins involved in the G1 to S phase transition. CLN1 and CLN2 are closely related genes with overlapping functions; both are expressed in late G1 phase when they associate with Cdc28p to activate its kinase activity. Accumulation of CLN1 and CLN2 mRNA in late G1 is dependent on two transcription factor complexes, MBF (Swi6p-Mbp1p) and SBF (Swi6p-Swi4p) which bind to MCB and SCB promoter elements respectively. In addition, Cln3p has been shown to activate CLN1 and CLN2 transcription while the G2 cyclins Clb1p, Clb2p, Clb3p, and Clb4p inhibit it. Pheromone induced cell cycle arrest is caused by the inhibition of the Cdc28p-Cln1p and Cdc28p-Cln2p complexes by the Far1 protein. An excellent review by Lew et al. describes cell cycle control in *S. cerevisiae* in detail.

Está anotado con los siguientes conceptos en Gene Ontology:

- [negative regulation of transposition, RNA-mediated \(IMP\)](#)
- [re-entry into mitotic cell cycle after pheromone arrest \(IGI\)](#)
- [regulation of cyclin-dependent protein kinase activity \(TAS\)](#)

Procesando todo el texto por ambas herramientas y buscando en IM en todo el texto que ofrece como resultado y en FastUMLS mirando sólo en los veinte primeros conceptos ofrecidos:

<b>FastUMLS</b>	<b>2 conceptos identificados</b>  GO:0000079: regulation of cyclin-dependent protein kinase activity (posición 69) GO:0000321: re-entry into mitotic cell cycle after pheromone arrest (posición 186)
<b>Interactive MetaMap</b>	<b>0 conceptos identificados.</b>

Es este caso no se han ofrecido los conceptos en las primeras 20 posiciones de los resultados de FastUMLS pero sí se han encontrado dos de los términos anotados manualmente dentro de los 200 primeros.

Puede decirse que el hecho de que IM trocee el texto origen en párrafos y frases le impide relacionar palabras muy alejadas en el texto, y que cuando éste es un resumen en el que las palabras alejadas suelen tener relación, esta estrategia tiene más inconvenientes que ventajas.

A pesar de que FastUMLS recupera de manera satisfactoria conceptos anotados por humanos expertos de manera manual, parece interesante comprobar si parte de los conceptos ofrecidos pudieran ser considerados como correctos por esos mismos anotadores. Este asunto se trata en cierta forma en la pruebas subjetivas mostradas más adelante en este documento.

## Comparación de resultados con Herramienta de PLN

Para poder realizar unas pruebas que permitieran comparar directamente los resultados ofrecidos por FastUMLS e *Interactive MetaMap*, se ha optado por procesar una de las descripciones de genes utilizadas anteriormente y utilizar como entrada para FastUMLS las diferentes frases que identifica IM.

IM ofrece para cada frase analizada un conjunto de conceptos candidatos y un “Meta Mapping” con los mejores conceptos relacionados con el texto tratado.

En este caso no se pone en duda la corrección de los resultados ofrecidos por una u otra herramienta, sino que se comparan directamente los conceptos ofrecidos, de forma que podamos comprender si la herramienta desarrollada ofrece resultados comparables con una de las herramientas de referencia para la extracción de conceptos anotados sobre textos biomédicos tal y como hemos referenciado a lo largo del trabajo [Zou 2003] y [French 2009].

El tipo y formato de los resultados ofrecidos por IM es el siguiente:

```
Processing 00000000.tx.1: RAP1 (Repressor Activator Protein) encodes an essential protein
involved in many diverse, some seemingly contradictory, processes in S. cerevisiae,
including telomere maintenance, transcriptional silencing (repression) of the silent mating
loci HML and HMR, and high level transcriptional activation of genes encoding ribosomal
proteins and glycolytic enzymes.

Phrase: "RAP1 Repressor Activator Protein"
>>>> Phrase
rap1 repressor activator protein
<<<<< Phrase
>>>>> Candidates
Meta Candidates (5):
  858 C0035147:Repressor Proteins {MSH, NDFRT} [Amino Acid, Peptide, or
Protein,Biologically Active Substance]
  812 C0033684:Protein (Proteins {CPM, LCH, LNC, MEDLINEPLUS, MSH, MTH, NCI, NDFRT, PSY,
RCD, RXNORM, SNOMEDCT, SNMI, AOD, FMA, SNM, UWDA, CSP, CCPSS, VANDF, NDDF, MEDCIN}) [Amino
Acid, Peptide, or Protein,Biologically Active Substance]
  812 C0202202:Protein NOS (Protein measurement {MTH, SNOMEDCT, SNMI, MDR, ICPC2P})
[Laboratory Procedure]
  645 C1336789:Repressor (Transcriptional Repressor {NCI}) [Amino Acid, Peptide, or
Protein,Biologically Active Substance]
  645 C1426113:RAP1 (TERF2IP gene {HUGO, MTH, OMIM}) [Gene or Genome]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (813):
  645 C1426113:RAP1 (TERF2IP gene {HUGO, MTH, OMIM}) [Gene or Genome]
  858 C0035147:Repressor Proteins {MSH, NDFRT} [Amino Acid, Peptide, or
Protein,Biologically Active Substance]
<<<<< Mappings
...

```

A continuación se exponen parte de los resultados del conjunto de frases analizadas sobre la descripción del Gen RAP1 con los conceptos ofrecidos por IM y los conceptos y posición de FastUMLS.

Frase	Interactive MetaMap	FastUMLS 10 conceptos	% Rec
"RAP1 Repressor Activator Protein"	<b>C1426113</b> : RAP1 <b>C0035147</b> : Repressor Proteins	C1426113 (1) C0035147 (5) C0905054 (2) RAP1 protein human	100,00%
"encodes"	<b>C1547699</b> : Encoding <b>C0679058</b> : encoding	C1547699 (2) C0679058 (1) C1707908: Encoder	100,00%
"an essential protein"	<b>C0205224</b> : Essential <b>C0202202</b> : Protein NOS <b>C0033684</b> : Protein	C0205224 (1)	33,33%
"some seemingly contradictory processes in S. cerevisiae"	<b>C1522240</b> : Process <b>C1951340</b> : Process Pharma <b>C1184743</b> : Process Bony	C1522240 (2) C1184743 (1) C0036025 (9) Cerevisiae	66,00%
"telomere maintenance,"	<b>C1155299</b> :	C1155299 (1)	100,00%

	Telomere Maintenance	C1336700 (2): Telomere Maintenance Gene	
"transcriptional silencing"	<b>C0040649:</b> Transcriptional <b>C0858952:</b> Silence	C0040649 (14)  C1156212 (1): transcriptional gene silencing	50,00%
"of the silent mating loci HML"	<b>C0443304:</b> Silent <b>C1260875:</b> mating <b>C1708726:</b> Locus <b>C1539104:</b> HML Gen	C1260875 (14) C1539104 (1) C0963478 (4): silent mating type information regulation.	50,00%
"ribosomal proteins"	<b>C0035552:</b> Ribosomal Proteins	C0035552 (1) C1622856 (2): Ribosomal Protein Activity C0035553 (4): Ribosome	100,00%

## Discusión de los Resultados

Los resultados obtenidos procesando la totalidad de las frases de descripción de los genes RAP1 y CLN2, indican que FastUMLS tiene un comportamiento más que aceptable en la mayoría de textos cortos analizados si lo comparamos con los ofrecidos por IM, y que sugiere los mismos conceptos (dentro de la lista de los 20 primeros) que InterActive MetaMap en más de un 60% de los casos.

Además y como ventaja inicial para FastUMLS, éste suele mostrar conceptos menos genéricos y más específicos a las palabras y términos analizados. Es por esto último por lo que podemos deducir que aparecen casos en los que FastUMLS no muestra alguno de los conceptos que sí ofrece IM.

FastUMLS tiende a penalizar los aciertos de palabras demasiado promiscuas en los conceptos, y es por eso que se justifican algunas ausencias, entendiéndose que no tiene que ser necesariamente una desventaja del proceso, sino quizás todo lo contrario.

En la frase "*of the silent mating loci HM*" FastUMLS muestra como primer concepto aquél que hace referencia al gen **HML**, y por el contrario omite el concepto *Silent*, como concepto cualitativo, debido a sobretodo a su generalidad. Pero FastUMLS en este caso, y como ocurre en muchos otros, ofrece en cuarto lugar un concepto que está plenamente relacionado con la idea del texto como es el proceso de regulación de información al que hace referencia C0963478 "*silent mating type information regulation*".

Debe tenerse en cuenta que FastUMLS hace uso de la última versión de los datos de 2009 e *Interactive MetaMap* utiliza la última versión de 2008, con algunas diferencias entre ellos.

## Pruebas Supervisadas

En esta fase de las pruebas han participado un total de cinco investigadores con publicaciones en los campos de la Biomedicina y la Biología Molecular, a los que se les ha pedido que seleccionen al menos cuatro artículos de los que ellos fueran autores o hubieran participado activamente, de forma que pudiera considerarse razonablemente que son expertos en el contenido tratado en cada uno de los artículos.

Una vez seleccionados los artículos, se extrajo el resumen de cada uno del campo "Abstract" desde la Web de PubMed/MEDLINE [Pubmed] y se procesó en el servicio Web de FastUMLS. Posteriormente se envió a cada investigador el Título de cada artículo junto con el enlace a los resultados con los conceptos ordenado por relevancia, con el siguiente cuestionario de cuatro preguntas:

**Para cada uno de los artículos y atendiendo sólo a los 20 primeros conceptos valore cada una de las siguientes preguntas:**

**1.- Los conceptos ofrecidos son correctos...**

- a) En un 100%.
- b) Entre 80% y 100%
- c) Entre 50% y 80%
- d) En menos de un 50%

**2.- Respecto a los conceptos ofrecidos, ¿se echa de menos algún concepto relacionado con el artículo?**

- a) No, ninguno.
- b) Sí, uno
- c) Sí, dos.
- d) Sí, más de dos.

**3.- Respecto a la especificidad de los conceptos ofrecidos... (un concepto demasiado general sería "enfermedad" en un artículo sobre "la enfermedad Alzheimer")**

- a) Son bastante específicos en su mayoría.
- b) En general hay conceptos específicos y genéricos a partes iguales.
- c) En general son demasiado genéricos.

**4.- La valoración general de los resultados es:**

- a) Muy buena.
- b) Buena, pero necesita mejorar.
- c) Media, necesita mejorar bastante.
- d) Mala.

Este cuestionario ha sido contestado por cada investigador para cada uno de los resúmenes de los artículos analizados, y además se han recibido comentarios escritos sobre mejoras o sugerencias de cambio.

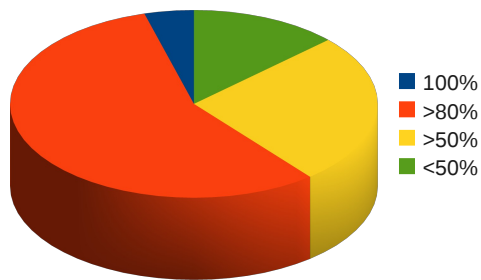
## Los investigadores y artículos tratados han sido:

<b>Virginia García de Yébenes, PhD</b> DNA Hypermutation and Cancer Group Centro Nacional de Investigaciones Oncológicas - CNIO Madrid	miR-181b negatively regulates activation-induced cytidine deaminase in B cells
	Polo-like kinase-1 is activated by aurora A to promote checkpoint recovery
	Expression of DNA damage checkpoint protein Hus1 in epithelial ovarian tumors correlates with prognostic markers
	Cleavage and degradation of Claspin during apoptosis by caspases and the proteasome
	Polo-like kinase-1 controls proteasome-dependent degradation of Claspin during checkpoint recover
<b>Emmanuelle Guillou, PhD</b> DNA Replication Group Centro Nacional de Investigaciones Oncológicas - CNIO Madrid	Mitochondrial morphology is controlled by large GTPases, such as Msp1p....
	Genomic DNA is packed in chromatin fibers that are organized in higher-order...
	Mitochondria are enveloped by two closely apposed boundary membranes with different ...
	Transmembrane segments of the dynamin Msp1p uncouple its functions in the control of mitochondrial morphology and genome maintenance.
<b>Raimundo Freire, PhD</b> Unidad de Investigación Hospital Universitario de Canarias La Laguna - Tenerife	Cell cycle-dependent processing of DNA lesions controls localization of Rad9 to sites of genotoxic stress
	Polo-like kinase-1 is activated by aurora A to promote checkpoint recovery
	Expression of DNA damage checkpoint protein Hus1 in epithelial ovarian tumors correlates with prognostic markers
	Cleavage and degradation of Claspin during apoptosis by caspases and the proteasome
	Polo-like kinase-1 controls proteasome-dependent degradation of Claspin during checkpoint recovery
<b>Veronique Smit, PhD</b> Unidad de Investigación Hospital Universitario de Canarias La Laguna - Tenerife	Polo-like kinase-1 is a target of the DNA damage checkpoint
	Negative growth regulation of SK-N-MC cells by bFGF defines a growth factor-sensitive point in G2
	ATR and Rad17 collaborate in modulating Rad9 localisation at sites of DNA damage
	Rapid PIKK-dependent release of Chk1 from chromatin promotes the DNA-damage checkpoint response
<b>Justo García de Yébenes, PhD</b> Jefe Unidad Enfermedades Neurodegenerativas Hospital Ramón y Cajal Madrid	NP7 protects from cell death induced by oxidative stress in neuronal and glial midbrain cultures from parkin null mice
	Gender differences and estrogen effects in parkin null mice
	Mortality, oxidative stress and tau accumulation during ageing in parkin null mice
	Plasma amyloid-beta, Abeta1-42, load is reduced by haemodialysis
	Drug-induced parkinsonism

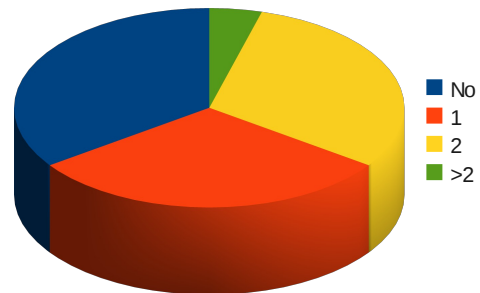
Y las respuestas al cuestionario son las siguientes:

Inv	Text	Corrección Conceptos				¿Faltan Conceptos?				¿Son específicos?			Valoración General			
		100%	>80%	>50%	<50%	No	1	2	>2	> 90%	>50%	<50%	Buena+	-Buena	Media	Mala
1	1		1			1				1			1			
	2		1			1				1			1			
	3		1			1				1			1			
	4		1			1				1			1			
	5		1			1				1			1			
2	1		1				1				1			1		
	2		1			1				1			1			
	3	1						1		1				1		
	4		1					1		1				1		
3	1			1		1				1				1		
	2		1				1				1			1		
	3			1			1			1					1	
	4			1			1				1			1		
	5				1		1			1					1	
4	1				1			1			1				1	
	2			1				1			1				1	
	3			1				1			1				1	
	4				1				1		1				1	
5	1			1		1					1			1		
	2		1				1				1				1	
	3		1					1			1				1	
	4		1					1			1				1	
	5		1					1			1				1	
%		4	57	26	13	35	30	30	4	48	52	0	26	35	39	0

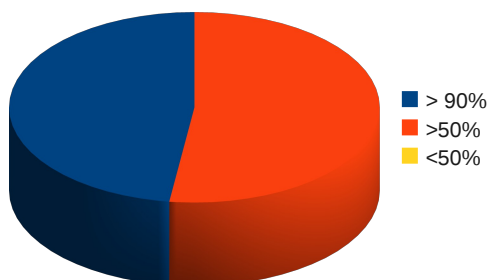
Corrección conceptos



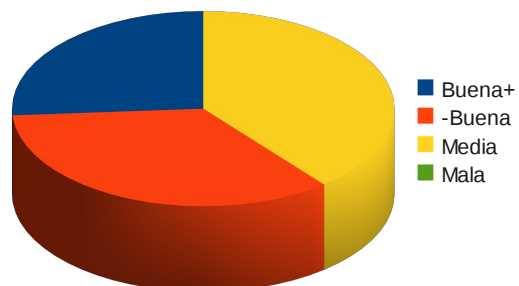
¿Faltan Conceptos?



¿Son Específicos?



Valoración General



Aunque de una manera subjetiva, con estas pruebas, se pretende evaluar la herramienta de acuerdo a los indicadores de porcentaje de recuperación de conceptos o “Recall”, respecto a la corrección de los conceptos mostrados o Precisión, y además en cuanto a lo específico o genérico de los conceptos, entendiendo como negativo que los resultados hagan referencia a términos demasiado generales.

<b>miR-181b negatively regulates activation-induced cytidine deaminase in B cells.</b>
Activated B cells reshape their primary antibody repertoire after antigen encounter by two molecular mechanisms somatic hypermutation SHM and class switch recombination CSR SHM and CSR are initiated by activation-induced cytidine deaminase AID through the deamination of cytosine residues on the immunoglobulin loci, which leads to the generation of DNA mutations or double-strand break intermediates. As a bystander effect, endogenous AID levels can also promote the generation of chromosome translocations, suggesting that the fine tuning of AID expression may be critical to restrict B cell lymphomagenesis. To determine whether microRNAs miRNA play a role in the regulation of AID expression, we performed a functional screening of an miRNA library and identified miRNAs that regulate CSR. One such miRNA, miR181b, impairs CSR when expressed in activated B cells, and results in the down-regulation of AID mRNA and protein levels. We found that the AID 3 untranslated region contains multiple putative binding sequences for miR181b and that these sequences can be directly targeted by miR-181b. Overall, our results provide evidence for a new regulatory mechanism that restricts AID activity and can therefore be relevant to prevent B cell malignant transformation

*Ejemplo de texto resumen procesado para su Revisión por los Investigadores*

**Recall** o Recuperación – R: Porcentaje de conceptos relacionados con el texto que se recuperan. Indica en qué medida la herramienta está mostrando todos los conceptos relacionados con el dominio del artículo.

**Precisión:** Porcentaje de los conceptos ofrecidos que tienen realmente que ver con el texto o palabras analizadas.

CUI	Matched strings	Terms related	Weight
<a href="#">C1842528</a> :L6465649	switch, recomb, csr, impair	(OMIM) MTHU002464: Impaired Ig class switch recombination (CSR) (OMIM) MTHU002417: Impaired Ig class-switch recombination (CSR)	8.957408
<a href="#">C1854499</a> :L6463015	somat, hypermut, shm	(OMIM) MTHU004532: Defective generation of somatic hypermutations (SHM)	8.815784
<a href="#">C1817371</a> :L6351237	immun, diversif, somat	(GO) <a href="#">GO:0002200</a> : somatic diversification of immune receptors	7.566369
<a href="#">C1511041</a> :L5361523	chromosom, transloc, balanc	(NCI) <a href="#">C6822</a> : Balanced Chromosomal Translocation (NCI) <a href="#">C6822</a> : Balanced Chromosomal Aberration/Rearrangement (NCI) <a href="#">C6822</a> : Balanced Chromosomal Alteration (NCI) <a href="#">C6822</a> : Balanced Chromosomal Rearrangement	7.250928
<a href="#">C1155229</a> :L2325408	humor, immun, respons	(GO) <a href="#">GO:0006959</a> : humoral immune response	7.165193
<a href="#">C0522274</a> :L0181112	humor, immun, defici	(NCI) <a href="#">C4799</a> : Deficiency of Humoral Immunity (ICD9CM) 279.0: Deficiency of humoral immunity (ICD9CM) 279.09: Other deficiency of humoral immunity (NCI) <a href="#">C4799</a> : B-Cell Deficiency (ICD9CM) 279.09: HUMORAL IMMUNITY DEF NEC	7.139249
<a href="#">C1709305</a> :L6057138	take, place	(NCI) <a href="#">C54069</a> : Occur (NCI) <a href="#">C54069</a> : Occurred (NCI) <a href="#">C54069</a> : Happen (NCI) <a href="#">C54069</a> : Happened (NCI) <a href="#">C54069</a> : Take Place (NCI) <a href="#">C54069</a> : Took Place	7.095090
<a href="#">C1421876</a> :L5080128	induc, cytidin, deaminas	(HUGO) 13203: AICDA gene (MTH) NOCODE: AICDA gene (OMIM) 605257: AID (OMIM) 605257: ACTIVATION-INDUCED CYTIDINE DEAMINASE (HUGO) 13203: activation-induced cytidine deaminase (HUGO) 13203: ARP2 (HUGO) 13203: CDA2 (HUGO) 13203: HIGM2 (OMIM) 605257: AICDA	7.090552
<a href="#">C2248928</a> :L7608644	class, switch, recomb	(GO) <a href="#">GO:0045190</a> : class switch recombination	7.042216

Ejemplo de Resultados Web de FastUMLS

## Discusión de los resultados

Los comentarios de los investigadores han sido inicialmente positivos y se han mostrado receptivos a utilizar una herramienta al estilo de FastUMLS para la anotación de conceptos, y casi todos han apuntado al interés la posibilidad de mejorar y enriquecer los resultados de las búsquedas que todos utilizan en *PubMed* o *Google Scholar* con los conceptos asociados a los documentos.

Las sugerencias de mejora pasan casi todas por tener en cuenta que hay conceptos que se repiten o son muy similares y por la posición que adoptan conceptos específicos a veces por debajo de conceptos más generales.

Un ejemplo de comentario al artículo 4 del investigador 1 es:

*"1) Aparecen con mayor jerarquía las excepciones que los conceptos generales. Este artículo es una revisión sobre los procesos de SHM y CSR y los dos primeros conceptos que aparecen son: "defective generation of SHM" y "Impaired Ig CSR". No son incorrectos, pero los generales, principales, aparecen con un ranking más bajo.*

*2) Al igual que en las búsquedas #1 y #2 aunque la mayoría de los conceptos son específicos, entre las primeras posiciones de resultados se encuentran algunos conceptos generales (los mismos que en las búsquedas #1 y #2)."*

La mayoría de investigadores ha apuntado que el porcentaje de conceptos correctos que esperarían estar relacionados con el artículo ("recall") mejora sensiblemente si se tienen en cuenta conceptos más allá de la posición número 20 dentro de la lista ordenada de conceptos.

Del total de los 23 artículos revisados por los investigadores:

### Respecto a la Corrección de los Conceptos:

- Más de un 60% de los análisis indican que se muestran conceptos adecuados o correctos en un porcentaje superior al 80%.
- Más de un 85% de los casos se piensa que más de la mitad de los conceptos ofrecidos son correctos.

### Respecto a la Recuperación de los Conceptos:

- En un 35% de los artículos no se han echado en falta conceptos relacionados con el tema tratado.
- El 95% de los casos indican que el número de conceptos a incorporar no es mayor que dos, mirando sólo los 20 primeros ofrecidos por la herramienta.
- En sólo un caso - 4% - se echan de menos más de dos conceptos en el mismo artículo.

### Respecto a lo Específico o Genérico de los conceptos:

- En un 48% de los artículos se valoran los conceptos como específicos en más de un 90% de todos los ofrecidos.

- En todos los casos se concluye que al menos la mitad de los conceptos son específicos al tema tratado.

**Valoración General del Resultado:**

- En más de un 60% de los artículos los investigadores piensan que la herramienta ofrece resultados buenos o más que buenos.
- Más de un 75% de los artículos sugieren que los resultados necesitan mejorar, y en casi un 40% que necesitan mejorar bastante.

## Líneas Futuras

Como se ha explicado en los resultados y a lo largo de este texto FastUMLS ha demostrado ser una herramienta capaz en la tarea de extracción automática de conceptos analizando textos biomédicos.

Se presentan a continuación algunas de las líneas de trabajo futuras para las que este trabajo, y otros, sirve de base.

### **Identificación de Grupos de Conceptos**

Usando técnicas de “*bi-clustering*” en Matrices SENT [Vazquez 2009] es capaz de agrupar conjunto de genes de una lista de ellos utilizados en experimentos con Micro-arrays. En ese caso la matriz está formada por dos dimensiones; la de genes y la de palabras. El concepto de “Característica Semántica” [Lee 1999] se utiliza para describir factores en la Factorización No-Negativa de Matrices (FNM o NMF del inglés “*Non-negative Matrix Factorization*”) que agrupan semánticamente palabras relacionadas.

Con estas características semánticas somos capaces de etiquetar el significado de una lista de genes, capturando los principales elementos, términos, que aparecen en los artículos relacionados con cada uno de los genes.

Como ya hemos comentado SENT utiliza palabras, y no conceptos formales como es el caso de los ofrecidos por UMLS.

Utilizar técnicas de bi-clustering [Prelic 2006] en combinación con los conceptos normalizados de Ontologías como UMLS es una de las líneas en el grupo en que surge este trabajo, enriqueciendo las búsquedas y los grupos con la expansión de los conceptos utilizando otros conceptos relacionados por los diferentes tipos de relaciones presentes en UMLS, tales como “es-parte” o “es-sinónimo”.

### **Enriquecimiento de Búsquedas:**

Una vez puesto a punto el proceso de anotación automática y mejorados algunos elementos (Ver las mejoras propuestas más adelante) una de las líneas interesantes basadas en FastUMLS, consiste en el procesamiento por lotes de grandes cantidades de literatura científica, como por ejemplo todos los resúmenes de artículos de PubMed o descripciones de experimentos masivos, para el etiquetado con conceptos en base de datos de cada uno de los textos.

Una vez hecha esta tarea la utilización posterior de dichas etiquetas para mejorar las búsquedas y aprovechar las relaciones jerárquicas entre los conceptos, parece ofrecer resultados prometedores a la hora de mejorar resultados de las búsquedas basadas en texto y en analizadores semánticos.

De esta forma búsquedas en las que una parte de la cadena sea el “Lóbulo Frontal” podrán tener en cuenta textos en los que aparezca el “Cerebro” sabiendo que hay una relación entre el primero como “es-una-parte” del segundo.

## **Mejoras al proceso actual**

Algunos de los avances al proceso ya han sido planteados en el trabajo y muchas de las veces han sido sugeridos por los investigadores usuarios y probadores del proceso.

### **Preprocesado de los datos y los pesos relativos:**

En muchas ocasiones existen elementos que distorsionan los resultados añadiendo conceptos en las primeras posiciones debido a la representatividad que tienen dentro de UMLS.

Por ejemplo, ocurre que en un artículo en el que aparecen las palabras “role” y “playing”, quizás en sentencias muy alejadas provocan la apariciones del concepto, “Role Play” como terapia usada en psicología, a pesar de que el contenido del artículo esté muy alejado de este tipo de asuntos. En este caso el problema viene de la baja cardinalidad que tiene “play” dentro de la base de datos: Sólo aparece relacionado con un concepto; el que ya se ha mencionado.

Identificar este tipo de problemas es una de las primeras mejoras a acometer para mejorar la precisión de FastUMLS.

El problema contrario parece fácilmente abordable. Existen palabras relacionadas con gran cantidad de conceptos y no aportan demasiado al significado del concepto, así que estas palabras podrían eliminarse si se cumplen determinadas condiciones iniciales.

En común también que las cadenas numéricas, por ejemplo 10, tiendan en general a introducir ruido más que a incrementar la precisión. En uno de los artículos analizados por uno de los investigadores se utilizaba una frase del estilo “Hace más de 10 años que...”, y el proceso unía la cadena “10” con el nombre de un gen para recuperar un concepto que consistía en el Gen con el calificador 10, y que resultaba completamente erróneo.

Curar, e incrementar el pre-procesado de los datos resulta necesario para eliminar ruido y mejorar los resultados del Proceso.

### **Tratamiento Semántico de los datos de entrada:**

Un análisis de los datos de entrada para eliminar o cambiar algunos de los términos mejoraría también los resultados. Al trabajar FastUMLS permutando las palabras de entrada, aunque de manera indirecta, no es capaz por ejemplo de detectar negaciones interpretándolas como afirmaciones; “*Non-cancerigenous derivative drugs*”, traería conceptos con los términos “*cancerigenous derivative drugs*”, que inicialmente no parecen correctos.

### **Eliminación de Duplicados:**

Esta mejora ha sido propuesta en varias ocasiones por los investigadores que han participado en las pruebas. Es relativamente sencillo de implementar y en éstos momentos ya se están dando los pasos para la eliminación de conceptos repetidos.

## Software y herramientas

Para la elaboración de FastUMLS y para la ejecución de pruebas se han utilizado software y servicios de terceros, en su mayoría con licencias de uso libre GPL.

### Entorno de Desarrollo:

- **Linux Ubuntu 8.10 Desktop Edition.**  
Sistema Operativo sobre el que se han desarrollado y ejecutado los procesos.  
<http://www.ubuntu.com/>
- **ADAPTIVE Communication Environment - ACE.**  
Framework de Desarrollo en C++ que permite desarrollar aplicaciones multiplataforma desde el mismo código (FastUMLS puede gracias a ACE ejecutarse en Linux, Windows, HP-UX o Solaris sin cambios).  
<http://www.cs.wustl.edu/~schmidt/ACE-overview.html>
- **GCC la Colección de Compiladores GNU**  
Compilador para C y C++  
<http://gcc.gnu.org/>
- **MySQL Client Libraries.**  
Librería de y código de conexión contra bases de datos MySQL.  
<http://www.mysql.com/>
- **Eclipse Platform CDT.**  
Entorno de Desarrollo Integrado para Java y C++.  
<http://www.eclipse.org/>
- **Ruby On Rails**  
Para Web services y Servicio Web.  
<http://rubyonrails.org/>

### Entorno de Producción:

- **Debian 4.1.2 para 64 bits**  
Sistema Operativo sobre el que se han implementado los diferentes procesos.  
Sobre servidor Intel Xeon a 3.40 GHz con 4 CPUs y 4GBytes de Memoria.  
<http://www.debian.org/>
- **MySQL 5.0.82 Community Server**  
Base de Datos Relacional con la Tablas de UMLS y las preprocesadas por FastUMLS.  
<http://www.mysql.com/>

- **Ruby on Rails y Web Services.**

Para página de servicio a usuarios desde Internet.

<http://rubyonrails.org/>

**Pruebas:**

- **FastUMLS Web**

Sitio Web para el procesamiento de texto y extracción de conceptos Biomédicos.

<http://bisaurin.dacya.ucm.es:8283/fastumls/analysis>

- **Interactive MetaMap**

Herramienta basada en Procesamiento de Lenguaje Natural para extracción de conceptos Biomédicos. Utilizada para comparar resultados.

<http://skr.nlm.nih.gov/interactive/metamap.shtml>

- **Saccharomyces Genome Database**

Sitio Web con información anotada sobre genes de la Levadura de la cerveza.

<http://www.yeastgenome.org>

## Referencias

### **[Agarwal 2008] “Briefings in Bioinformatics 2008”**

Pankaj Agarwal and David B. Searls

Briefings in Bioinformatics Special Issue: Database Integration in Life Sciences – 2008

### **[Aronson 2001] “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program”**

AR Aronson

Proc. AMIA Symposium – 2001.

### **[Aronson 2008] “Methodology for Creating UMLS Content Views Appropriate for Biomedical Natural Language Processing”**

AR. Aronson, PhD, James G. Mork, MSc, Aurélie Névéol, PhD, Sonya E. Shooshan, MLS, and Dina Demner-Fushman, MD, PhD

### **[Ashburner 2000] “Gene Ontology: tool for the unification of biology”**

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, S Lewis, JC. Matese, JE. Richardson, M Ringwald, GM. Rubin & Sherlock

*The Gene Ontology Consortium - Nature Genetics* 25, 25 – 29 – 2000

### **[Bodenreider 2004] “The Unified Medical Language System (UMLS): integrating biomedical terminology”**

Olivier Bodenreider

Nucleid Acid Research D267-D270Vol 32 - 2004

### **[Divita 2004] “Failure analysis of MetaMap Transfer (MMTx)”**

Divita G, Tse T, Roth L.

MedInfo – 2004

### **[French 2009] “Application and evaluation of automated semantic annotation of gene expression experiments”**

Leon French, Suzane Lane, Tamryn Law, Lydia Xu and Paul Pavlidis

Bioinformatics – 2009

### **[Greenes 2007] “Clinical decision support: The road ahead”**

RA. Greenes

Boston – Elsevier Academic Press – 2007

**[GO]The Gene Ontology Consortium Wiki**

[http://wiki.geneontology.org/index.php/Main\\_Page](http://wiki.geneontology.org/index.php/Main_Page)

**[Google Scholar] Google Scholar o Google Académico.**

<http://scholar.google.es/>

**[Kaminski 2000] “Bioinformatics: A user's perspective”**

N. Kaminski

American Journal of Respiratory Cell and Molecular Biology – 2000

**[I-METAMAP] Interactive MetaMap**

Es necesario utilizar un usuario y clave suministrados después de aceptar las condiciones de uso y licencia.

<http://skr.nlm.nih.gov/interactive/metamap.shtml>

**[Lee 1999] Learning the parts of objects by non-negative matrix factorization**

Daniel D. Lee & H. Sebastian Seung

Nature 401 – 1999

**[McEntyre 2001] “PubMed: Bridging the information gap”**

Johanna McEntyre and David Lipman

Canadian Medical Association Journal – 2001

**[Nadkarni 2001] “UMLS Concept Indexing for Production Databases”**

Prakash Nadkarni, MD, Roland Chen, MD and Cynthia Brandt, MD, MPH

Journal of the American Medical Informatics Association 8:80-91 - 2001

**[Porter 1980] “An Algorithm for suffix stripping”**

M.F. Porter

Program: Electronic Library & Information System – 1980 – Rep. 2006

**[Prelic 2006] “A systematic comparison and evaluation of biclustering methods for gene expression data”**

Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E.

Bioinformatics – 2006

**[PubMed] National Center for Biotechnology Information (NCBI)**

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

**[Smith 2007] “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration”**

B Smith, M Ashburner, C Rosse, J Bard, W Bug, W Ceusters, LJ Goldberg, K Eilbeck, A Ireland, CJ Mungall, The OBI Consortium, N Leontis, P Rocca-Serra, A Ruttenberg, SA Sansone, RH Scheuermann, N Shah, PL Whetzel16 & S Lewis

Nature Biotechnology 25 – 2007

**[Stevens 2007] “Using OWL to model biological knowledge”** R Stevens, M Egaña, K Wolstencroft, Ulrike Sattlera, Nick Drummond, Matthew Horridge and Alan Rectora

International Journal of Human-Computer Studies - Elsevier – 2007

**[Vazquez 2009] “SENT: semantic features in text”**

Miguel Vazquez, Pedro Carmona-Saez, Ruben Nogales-Cadenas, Monica Chagoyen, Francisco Tirado, Jose Maria Carazo and Alberto Pascual-Montano

Nucleic Acids Research - doi:10.1093/nar/gkp392 – 2009.

**[Zou 2003] “IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing”**

Qinghua Zou, MSc, Wesley W. Chu, PhD, Craig Morioka, PhD, Gregory H. Leazer, PhD, and Hooshang Kangarloo, MD

American Medical informatics Association – AMIA Annual Symp – 2003

## **Agradecimientos**

A Mariana, a Rubén, a Pedro y a Alberto.

A Raimundo, a Veronique, a Justo y a Manu.

A Virginia.