Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 27-34 recibido 31-01-08, aceptado 03-03-08

Detección de fármacos genéricos en textos biomédicos

Detecting generic drugs in biomedical texts

Isabel Segura Bedmar
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés, Madrid
isegura@inf.uc3m.es

Paloma Martínez
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés, Madrid
pmf@inf.uc3m.es

Doaa Samy
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés, Madrid
dsamy@inf.uc3m.es

Resumen: Este trabajo presenta un sistema para el reconocimiento y clasificación de nombres genéricos de fármacos en textos biomédicos[1]. El sistema combina información del Metatesauro UMLS[2] y reglas de nomenclatura para fármacos genéricos, recomendadas por el consejo "United States Adoptated Names" (USAN)[3], que permiten la clasificación de los fármacos en familias farmacológicas. La hipótesis de partida es que las reglas USAN son capaces de detectar posibles candidatos de fármacos que no están incluidos en UMLS (versión 2007AC), aumentando la cobertura del sistema. El sistema consigue un 100% de precisión y un 97% de cobertura usando sólo UMLS sobre una colección de 1481 resúmenes de artículos científicos de PubMed. La combinación de las reglas USAN con UMLS mejoran ligeramente la cobertura del sistema.
Palabras clave: Reconocimiento de entidades biomédicas, Fármacos Genéricos, UMLS

Abstract: This paper presents a system for drug name recognition and clasification in biomedical texts. The system combines information from UMLS Metathesaurus and nomenclatura rules for generic drugs, recommended by United States Adoptated Names (USAN), that allow the classification of the drugs in pharmacologic families. The initial hypothesis is that rules are able to detect possible candidates of drug names which are not included in the UMLS database (version 2007AC), increasing, in this way, the coverage of the system. The system achieves a 100% precision and 97% recall using UMLS only. The combination of the USAN rules and UMLS slightly improves the coverage of the system.
Keywords: Biomedical Named Entities, Generic Drugs, UMLS.

1 Introducción

Este trabajo es un primer paso en el desarrollo de un sistema que permita la extracción automática de interacciones farmacológicas en textos biomédicos. Una interacción ocurre cuando los efectos de un fármaco se modifican por la presencia de otro fármaco, o bien de un alimento, una bebida o algún agente químico ambiental (Stockley, 2004).

Las consecuencias pueden ser perjudiciales si la interacción causa un aumento de la toxicidad del fármaco. Por ejemplo, los pacientes que reciben *warfarina* pueden comenzar a sangrar si se les administra *azapropazona* o *fenilbutazona* sin disminuir la dosis de *warfarina*. Del mismo modo, la disminución de la eficacia de un fármaco causada por una interacción puede ser igual de peligrosa: si a los pacientes que reciben *warfarina* se les administra *rifampicina*, necesitaran más cantidad de aquélla para mantener una anticoagulación adecuada. Sin

[1] Este trabajo ha sido parcialmente financiado por los proyectos FIT-350300-2007-75 (Interoperabilidad basada en semántica para la Sanidad Electrónica) y TIN2007-67407-C03-01 (BRAVO: Búsqueda de respuestas avanzada multimodal y multilingüe).
[2] http://www.nlm.nih.gov/research/umls/
[3] http://www.ama-assn.org/ama/pub/category/2956.html

embargo, en determinadas ocasiones el uso combinado de medicamentos puede ser beneficioso. La combinación de fármacos antihipertensivos y diuréticos logran unos efectos antihipertensores que no se obtendrían con la administración de uno u otro fármaco por separado *(Stockley, 2004)*.

Cuantos más fármacos toma un paciente, mayor es la probabilidad de producirse una interacción adversa. En un estudio hospitalario se halló que el porcentaje era del 7% entre aquellos pacientes que tomaban entre 6 y 10 fármacos, pero aumentaba en un 40% en aquellos que ingerían entre 16 y 20 fármacos, lo que representa un aumento desproporcionado (Smith et al., 1969).

Investigadores y profesionales de la salud utilizan distintos recursos como bases de datos online y herramientas[4,5] para identificar y prevenir las interacciones farmacológicas. Sin embargo, la literatura biomédica es el mejor sistema para estar al día en lo que se refiere a la información sobre nuevas interacciones.

Los últimos avances en biomedicina han provocado un crecimiento vertiginoso del número de publicaciones científicas. PubMed[6], un buscador online de artículos de la revista MedLine, tiene más de 16 millones de resúmenes. Investigadores y profesionales de la salud están desbordados ante tal avalancha de información.

Por este motivo, es imprescindible el desarrollo de sistemas que faciliten la extracción de conocimiento y un acceso eficiente a la información en el dominio de la biomedicina. El uso de recursos y tecnologías de procesamiento de lenguaje natural puede contribuir a ello.

El reconocimiento y clasificación de los términos biomédicos es una fase crucial en el desarrollo de este tipo de sistemas. Es imposible comprender un artículo sin una precisa identificación de sus términos (genes, proteínas, principios activos, compuestos químicos, etc.).

La detección de nombres de fármacos genéricos es una tarea compleja debido a las dificultades que implica el procesamiento del texto farmacológico. Nuevos fármacos se introducen diariamente mientras que otros se retiran. Los recursos terminológicos, aunque se modificados frecuentemente, no pueden seguir el paso acelerado de esta terminología en constante cambio. Así, los sistemas capaces de detectar de forma automática nuevos fármacos pueden contribuir a la actualización automática de sus bases de conocimiento.

El sistema presentado en este artículo persigue el reconocimiento y clasificación de nombres genéricos de fármacos, combinando información de UMLS y un módulo que implementa las reglas recomendadas por el consejo USAN para la denominación de sustancias farmacológicas. Esta fase es un paso previo e imprescindible para la extracción automática de las interacciones farmacológicas en la literatura biomédica.

La combinación de ambos recursos obtiene una precisión y cobertura elevada. UMLS garantiza la precisión, mientras que las reglas amplían la cobertura del dominio detectando nuevos nombres de fármacos que aún no han sido registrados en UMLS.

Además, las reglas permiten una clasificación más específica de los fármacos en familias farmacológicos, que ULMS no es capaz de aportar. Consideramos que la familia de un fármaco puede ser una pista valiosa a la hora de detectar interacciones farmacológicas en textos biomédicos. Los fármacos de una misma familia comparten una estructura química base, y por este motivo, si es conocida la interacción de un determinado fármaco, es bastante probable que otro fármaco de la misma familia presenten la misma interacción.

El artículo está organizado como sigue: la sección 2 es una revisión de los trabajos en el reconocimiento de entidades biomédicas. La sección 3 describe brevemente los principales recursos de información utilizados en el sistema: UMLS y las reglas USAN. La sección 4 proporciona una descripción de la arquitectura del sistema y el corpus utilizado. La evaluación se presenta en la sección 5. Finalmente, la sección 6 incluye algunas conclusiones y el trabajo futuro.

2 Trabajos relacionados

La identificación de genes, proteínas, compuestos químicos, fármacos y enfermedades, etc., es crucial para facilitar la recuperación de información y la identificación de relaciones entre esas entidades, como por ejemplo, las interacciones entre fármacos.

[4] http://www.micromedex.com/products/
[5] http://www.ashp.org/ahfs/index.cfm
[6] http://www.ncbi.nlm.nih.gov/sites/entrez/

El reconocimiento de entidades intenta encontrar términos de interés en el texto y clasificarlos dentro de categorías predefinidas como genes, compuestos químicos, fármacos, etc. El problema consiste en determinar dónde empieza y termina cada término, y la asignación de la clase correcta.

Muchos trabajos se han centrado en la identificación de genes (Tanabe y Wilbur, 2002) y proteínas (Fukuda et al., 1998). Menor atención ha recibido la detección de otro tipo de entidades como las sustancias químicas (Wilbur et al., 1999), fármacos (Rindflesch et al., 2000) o enfermedades (Friedman et al., 2004).

Se han empleado diferentes enfoques para tratar el problema del reconocimiento de entidades biomédicas: reglas, diccionarios, aprendizaje automático, métodos estadísticos, y una combinación de las distintas técnicas. Los métodos basados en diccionarios utilizan recursos terminológicos para localizar las ocurrencias de los términos en el texto. Su principal desventaja es que no son capaces de tratar adecuadamente la variabilidad terminológica. Normalmente, un mismo concepto puede recibir distintos nombres, y los diccionarios, en numerosas ocasiones, no recogen esta variabilidad.

(Hirschman et al, 2002) utiliza patrones para localizar genes en una lista extensa obtenida de la base de datos FlyBase. Muchos nombres de genes comparten su representación léxica con palabras comunes en el idioma inglés (ej: *an, by, can, for*). Esta homonimia es la responsable de la baja precisión del sistema: un 2% en artículos completos y un 7% en resúmenes. La cobertura varía de 31% en resúmenes a un 84% en artículos completos.

En (Tsuruoka y Tsujii, 2003) se describe un método para el emparejamiento aproximado de cadenas en un diccionario de proteínas. Además, este método utilizaba un clasificador Bayesiano entrenado sobre el corpus GENIA[7], para filtrar los falsos positivos. Este filtrado mejora la precisión (73.5%), al excluir ciertos términos detectados como proteínas según el diccionario, pero que realmente no lo son en el texto. El sistema consigue una cobertura del 67.2%.

El principal enfoque de los sistemas basados en reglas consiste en el desarrollo de heurísticas o gramáticas que describan las estructuras comunes de los nombres de determinadas entidades mediante el uso de pistas léxicas y ortográficas, aunque también se suele utilizar información morfosintáctica. Una de sus principales desventajas es el elevado coste de tiempo y esfuerzo que implica el desarrollo de las reglas. Además, su adaptación para el reconocimiento de otro tipo de entidades es compleja. La combinación de elementos internos tales como afijos, raíces, letras griegas y latinas se emplea para describir la formación de patrones de términos mediante una gramática en el trabajo (Ananiadou, 1994).

El sistema PROPER, desarrollado por (Fukuda et al., 1998), utiliza patrones léxicos y elementos ortográficos para la detección de nombres de proteínas, consiguiendo en un pequeño experimento una precisión del 94.7% y una cobertura del 98.8%. El sistema PASTA utiliza una gramática libre de contexto para el reconocimiento de proteínas. Las reglas están basadas en propiedades léxicas y morfológicas de los términos del dominio. El sistema consigue un 84% de precisión y un 82% de cobertura en el reconocimiento de 12 clases de proteínas (Gaizauskas et al., 2003). En el trabajo de (Narayanaswamy et al., 2003) se combina el uso de raíces y sufijos típicos en el dominio químico, con información contextual, es decir, información sobre las palabras que rodean la entidad. También hay trabajos de adaptación de reconocedores de entidades de carácter general com el presentado en (Hobbs, 2002) para detección de nombres de proteínas.

Otros enfoques combinan el uso de diccionario y reglas para mitigar el problema de la variabilidad terminológica, y conseguir así una mayor cobertura. (Chiang y Yu, 2003) proponen un sistema robusto de reconocimiento de términos basado en reglas y en la ontología Gene[8]. Las reglas consideran las posibles variaciones multipalabra, generadas por las permutaciones y por la inserción o eliminación de palabra individuales.

Menor es el número de los sistemas que han utilizado aprendizaje supervisado, debido principalmente a la carencia de corpus etiquetados en el dominio biomédico. A continuación, se presentan algunos de estos sistemas basados en aprendizaje automático.

En (Zhan et al., 2004) se adaptó un modelo oculto de Markov para el reconocimiento de entidades y abreviaturas en el dominio

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[8] http://www.geneontology.org/

biomédico, mediante el uso de elementos ortográficos, morfológicos, morfosintácticos y semánticos. (Collier y Takeuchi, 2004) utilizan el clasificador Support Vector Machines (SVM) para detectar entidades biomédicas. Los elementos utilizados fueron ortográficos y etiquetas morfosintácticas. Los experimentos demostraron que el uso de información morfosintáctica provocaba un ligero descenso en los resultados.

En (Lee et al, 2004), el reconocimiento se divide en dos fases: identificación y clasificación. Esta división permite una selección más apropiada de los elementos utilizados para el entrenamiento del algoritmo SVM en cada una de las fases.

El sistema descrito en este artículo combina el uso de reglas y diccionario. Las reglas están basadas en las recomendaciones del consejo USAN para nominar sustancias farmacológicas. Además, la utilización de estándares oficiales, como es el caso de las reglas USAN, garantiza cierta precisión comparada con la que podría obtenerse al aplicar simples heurísticas.

3 Recursos específicos del sistema

El sistema utiliza dos fuentes de información para identificar y clasificar los nombres de fármacos en textos biomédicos: el Metatesauro UMLS y las recomendaciones del consejo USAN para el nombrado de fármacos genéricos. Ambos se describen a continuación.

3.1 UMLS Knowledge Sources (UMLSKS)

El Sistema de Lenguaje Médico Unificado (UMLS) es una base de datos de conocimiento que integra varios recursos. Uno de sus principales propósitos es facilitar el desarrollo de sistemas automáticos para el procesamiento lenguaje natural en el dominio de la biomedicina. Tres son los recursos principales de UMLS: el Metatesauro, la red semántica y el SPECIALIST Lexicón.

El Metatesauro solventa en gran medida el problema de la variabilidad terminológica, debido a que integra información de más de 60 vocabularios y clasificaciones biomédicas. La organización del Metatesauro está basada en conceptos. Un concepto agrupa los posibles nombres que puede tomar un mismo significado en la literatura médica. En el sistema aquí presentado, el Metatesauro UMLS permite la identificación de los nombres de fármacos en el texto.

La red semántica consta de 135 tipos semánticos y 54 relaciones que representan relaciones importantes en el dominio de la biomedicina. La Figura 1 muestra parte de la red semántica. Cada concepto de UMLS se clasifica por al menos un tipo semántico. Debido a su extenso alcance, la red semántica permite la categorización de un amplio rango de terminología, lo que favorece el desarrollo de sistemas para el procesamiento automático del lenguaje natural en múltiples dominios biomédicos.

Sin embargo, en lo que se refiere al dominio farmacológico, esta categorización es insuficiente. En UMLS, los fármacos genéricos se clasifican en “Pharmacological Substances” o “Antibiotics”. El tipo “Clinical Drugs” se refiere a marcas comerciales, y queda fuera del alcance de nuestro estudio.

Mientras que los antibióticos se clasifican en el tipo “Antibiotics”, para el resto de familias farmacológicas (analgésicos, antivirales, anticoagulantes, antiinflamatorios, etc), UMLS proporciona una clasificación demasiado general, al clasificarlos como “Pharmacologic Substance”, sin hacer distinción alguna entre las distintas familias.

El tercer recurso de UMLS, SPECIALIST Lexicón está formado por numerosos términos biomédicos y contiene información sintáctica, morfológica y ortográfica.

Es posible acceder a estos recursos de tres formas distintas: a través de un servidor cliente utilizando un navegador estándar, mediante un programa que utilice el API UMLSKS, o a través de una interfaz TCP/IP. También es posible trabajar con una copia local de los recursos UMLS, distribuida gratuitamente por la National Library Medical (NLM)[9] de Estados Unidos. En la arquitectura aquí descrita se implementó un programa JAVA que embebía el API UMLSKS para acceder a la información en el servidor remoto.

3.2 Reglas de nombrado recomendadas por el consejo USAN.

Un fármaco tiene tres nombres: uno químico basado en su estructura, uno genérico (no propietario) que es el nombre oficial del fármaco durante su existencia, y la marca

[9] http://www.nlm.nih.gov/

comercial que es el nombre dado por la compañía farmacéutica que lo comercializa.

La selección de un nombre para un nuevo fármaco es un proceso complejo. En Estados Unidos, el consejo *U.S. Adopted Name* (USAN) es la institución responsable de la creación y asignación de un nombre genérico a un nuevo fármaco. En la selección de un nombre, se consideran los siguientes aspectos: la seguridad del paciente, la facilidad de pronunciación, la ausencia de conflictos con marcas comerciales y la utilidad para los profesionales de la salud.

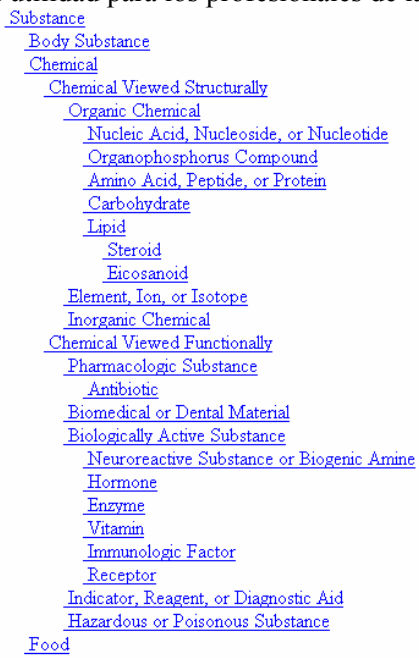

Figura 1 Un subconjunto de la Red Semántica de UMLS

Las prácticas actuales para nombrar fármacos recaen en el uso de afijos. Estos afijos clasifican los fármacos dependiendo de su estructura química, indicación o mecanismo de acción. Por ejemplo, el nombre de un analgésico podría contener alguno de los siguientes afijos:–*adol, -adol-, -butazone, -fenine, -eridine* y *–fentanil.*

En este trabajo, la clasificación de los fármacos se ha basado en los afijos recomendados por USAN[10]. La lista utilizada no es exhaustiva, debido a que no incluye ni todos los afijos aprobados por el consejo USAN, ni los recomendados por otras organizaciones. La Tabla 1 muestra algunos de los sufijos empleados en la clasificación.

[10] http://www.ama-assn.org/ama/pub/category/4782.html

La categorización en familias farmacológicas proporcionada por los afijos es más específica y detallada que la proporcionada por los tipos semánticos de UMLS. Además, los afijos permiten identificar nombres de fármacos que aún no han sido registrados en el Metatesauro UMLS.

Afijos	Definición
-ast	antiasthmatics/antiallergics
-cromil	antiallergics (cromoglicic). Ej: ***nedocromil***
-atadine	tricyclic antiasthmatics. Ej: ***olopatadine***
-tibant	antiasthmatics (bradykinin antagonists). Ej: ***icatibant***
-adol, -adol-	analgesics (mixed opiate receptor agonists/antagonists). Ej: ***tazadolen***
-butazone	anti-inflammatory analgesics. Ej: ***mofebutazone***
-eridine	analgesics (meperidine). Ej: ***anileridine***
-fenine	analgesics (fenamic). Ej: ***floctafenine***
-fentanil	narcotic analgesics. Ej: ***alfentanil***
-adox	antibacterials (quinoline dioxide). Ej: **carb*adox***
-ezolid	oxazolidinone antibacterials Ej: ***eperezolid***
-mulin	antibacterials (pleuromulin) Ej: ***retapamulin***
-penem	antibacterial antibiotics, Ej: ***tomopenem***
-oxacin	antibacterials (quinolone). Ej: ***difloxacin***
-planin	antibacterials (*Actinoplane*) Ej: ***mideplanin***
-prim	Antibacterials (trimethoprim type). Ej: ***ormetoprim***
-pristin	Antibacterials (pristinamycin) Ej: ***quinupristin***
-arol	anticoagulants (dicumarol). Ej: ***dicumarol***
-irudin	anticoagulants (hirudin). Ej: ***desirudin***
-rubicin	antineoplastic antibiotics (daunorubicin) Ej: ***esorubicin***
-fungin	antifungal antibiotics Ej. ***kalafungin***

Tabla 1: Algunos afijos empleados por USAN

4 *Descripción del sistema*

Se ha trabajado con una colección de 1481 resúmenes de artículos científicos de PubMed recuperada mediante búsquedas de los nombres

de familias farmacológicas, tales como “antiallergics”, “antiasthmatics”, “analgesics”, “antibacterials”, “anticoagulants”, etc. Esta colección se obtuvo mediante un Web Crawler implementado para la recuperación de los resúmenes.

La arquitectura del sistema (Figura 2) consta de tres módulos que se ejecutan de forma secuencial: (1) un módulo encargado del procesamiento de los resúmenes, (2) un módulo que identifica los términos que son fármacos, y por último, (3) el módulo responsable de la clasificación y de detectar nuevos fármacos que aún no han sido registrados en UMLS. Para cada uno de los resúmenes de la colección, cada módulo produce como salida un fichero XML con la información obtenida por él.

En primer lugar, los resúmenes se dividen en oraciones, se identifican los tokens y se analizan morfosintácticamente. Este módulo utiliza los procesos *Sentence Splitter*, *Tokenizer* y *POS tagger* de la infraestructura GATE[11].

El análisis morfosintáctico es necesario para identificar aquellos tokens cuya categoría morfosintáctica es nombre (común, propio o plural). A continuación, cada uno de estos nombres se busca en WordNet para descartar aquellos nombres que no son específicos del dominio biomédico, debido a que WordNet es un lexicón de carácter general. La lista inicial de candidatos está formada por aquellos nombres no encontrados en WordNet.

El segundo módulo busca en el Metatesauro de UMLS cada uno de los términos que no han sido encontrados en WordNet. Esta búsqueda es implementada utilizando el API de Java que proporciona UMLSKS y que permite consultar información en su servidor remoto.

El servidor devuelve un fichero XML con los resultados de la búsqueda. Si se ha encontrado uno o más conceptos, el módulo trata la respuesta y localiza sus posibles tipos semánticos. Si ninguno de ellos se corresponda con “Pharmacological Substance” o “Antibiotics” entonces el término pertenece a otro tipo de entidades (genes, proteínas, etc.). Aunque estas entidades están fuera del alcance del presente estudio, la información relativa a sus tipos semánticos, así como el nombre del concepto, idioma, recurso de información origen, y su identificación dentro de UMLS, queda registrada en el fichero XML que produce el módulo como salida. Si por el contrario, alguno de los tipos semánticos es “Pharmacologic Substance” o “Antibiotic”, el término se etiqueta como fármaco, junto el resto de la información obtenida de UMLS.

Los términos que no se encuentran en UMLS, se etiquetan como candidatos a nuevos fármacos no registrados en UMLS.

Por último, el modulo que implementa las recomendaciones del consejo USAN es el responsable de clasificar los términos etiquetados como fármacos por el modulo anterior. Para cada uno de los términos, el modulo devuelve la lista de los afijos que están contenidos dentro del nombre, consiguiendo así, la lista de sus posibles familias farmacológicas.

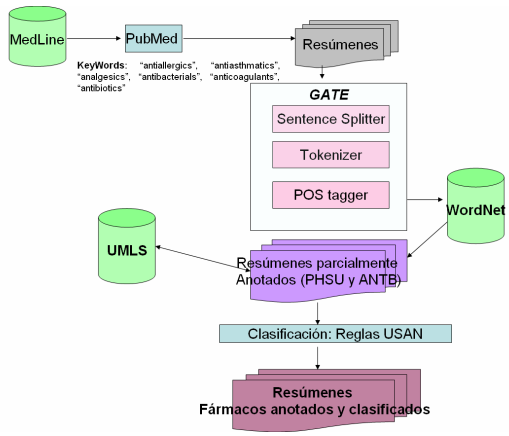

Figura 2. Arquitectura del sistema

Algunos afijos son demasiado ambiguos, tales como: -ac, -vin-,-vir-, -vin, -mab-, -kin, -glil-, -dil, -sal- etc. Dichos afijos podrían disminuir la precisión del sistema, clasificando términos en familias incorrectas. Por este motivo, en la implementación del módulo se decidió prescindir de los afijos con menos de tres letras. Claramente, la clasificación no es exhaustiva, debido a la eliminación de estos afijos ambiguos, y al hecho de que la lista considerada inicialmente no era completa. Por otro lado, con el objeto de detectar posibles candidatos de nuevos fármacos que aún no han sido registrados en el Metatesauro, el módulo procesa el conjunto de términos que no fueron encontrados en UMLS. Como se analizará en el siguiente apartado, el número de nuevos candidatos detectados exclusivamente por las reglas es muy pequeño.

[11] http://www.gate.ac.uk/

5 Evaluación del sistema

Una vez procesados los 1481 resúmenes y descartados los nombres de dominio general, es decir, aquellos que fueron encontrados en WodNet, la lista inicial de candidatos está formada por 10.743 tokens.

Cada uno de estos términos se busca en el metatesauro de UMLS. Un 10.5% de ellos (1.129) están registrados en el Metatesauro, pero ninguno de sus tipos semánticos es "Pharmacologic Substances" o "Antibiotics". Es decir, estos términos pertenecen a otros tipos semánticos como "Organic Chemical", "Lipid2, "Carbohydrate", etc., Como se comentó anteriormente, este subconjunto está fuera del alcance del presente estudio.

El 75.4% (8.103) de los 10.743 candidatos iniciales se corresponden con sustancias farmacológicas o antibióticos.

El módulo que implementa las reglas USAN consigue clasificar un 35% (2.893) de ellos. La Tabla 2 muestra parte de la distribución de familias farmacológicas en la colección de resúmenes.

Familia	Afijos	% (num)
Antineoplastics	-abine, -antrone, -bulin, -platin, -rubicin, -taxel, -tinib, -tecan, -trexate, -vudine	7% (205)
Anticoagulants	-arol-, -grel-	1,3%(37)
Antihistaminics	-tadine, -astine	1,5%(44)
antiasthmatics or antiallergics	-azoline, -cromil	2,1%(61)
Anxiolytic sedatives	-azenil, -azepam, -bamete, -peridone, -perone	0,8%(24)
Antibacterials	-ezolid, -mulin, -oxacin, -penem, -planin, -prim, -pristin	5%(146)
Antifungals	-conazole, -fungin	1,8%(53)
Antivirals	-cavir, -ciclovir, -navir, -vudine, -virenz,	4,7%(137)
Anti-inflammatory	-bufen, -butazone, -icam, -nidap, -profen,	4,9%(141)
Immunomodulators	-imod, -leukin	5,3%(154)
Antidiabetics	-glinide, -glitazone	0,7%(22)
Vasodilators	-dipine, -pamil	2,4%(71)
Analgesics	-adol, -butazone, -coxib -eridine, -fentanil	3,9%(115)

Tabla 2. Distribución de las familias farmacológicas en el corpus

UMLS no detectó ningún concepto para el 14% (1.511) de los candidatos iniciales (10.743). Aunque UMLS es un recurso actualizado frecuentemente y con una elevada cobertura en el dominio de la farmacología, pensamos que las reglas USAN podrían detectar fármacos que aún no han sido registrados en el metatesauro. Por este motivo, el módulo de clasificación se ejecutó sobre este conjunto, detectándose 102 nuevos candidatos. Un experto del dominio evaluó manualmente el conjunto de candidatos concluyendo que sólo 82 de estos candidatos eran realmente fármacos no incluidos en UMLS (versión 2007AC). Algunos ejemplos de estos fármacos son: spirad**olene**, mide**planin**, efe**pristin**, tomo**penem**.

Del resto de candidatos, 579 se correspondían con entidades del dominio general tales como organizaciones, nombres de personas, etc. Esto se debe a que los resúmenes, además de contener el título del artículo, también contenían información sobre los autores y su afiliación que no se había filtrado previamente. Los restantes 830 son términos del dominio de la biomedicina que no están registrados en UMLS, tales como *nonherbal, suboptimal, thromboprophylaxis, interpatient, coadministration,* etc.

Finalmente, los resultados globales de la evaluación se muestran en la Tabla 3. El sistema consigue una cobertura del 97% y una precisión del 100% si se utiliza únicamente información de UMLS. La combinación de UMLS y las reglas USAN aumentan ligeramente la cobertura, pero disminuye la precisión del sistema.

	Cobertura	***Precisión***
UMLS	97%	100%
UMLS + Rules	99.8%	99,3%

Tabla 3. Resultados del sistema

6 Conclusiones

La implementación de las reglas USAN puede mejorar la detección de nuevos fármacos aún no registrados en el Metatesauro UMLS. Sin embargo, los resultados demuestran que la mejora es realmente pequeña. Por esta razón, es lógico concluir que UMLS tiene una elevada cobertura en el dominio de la farmacología.

Por otro lado, la categorización aportada por UMLS en lo que se refiere a los fármacos es insuficiente a la hora de desarrollar sistemas automáticos para la extracción automática de

información. Las reglas USAN pueden contribuir a completar la clasificación de UMLS. Conocer la clase o familia de un determinado fármaco es una valiosa pista a la hora de determinar la presencia real de una interacción.

Este enfoque preliminar es el primer paso hacia un sistema de extracción de información en el campo de la farmacología. Ampliar la cobertura de la clasificación gracias a la inclusión de un mayor número de afijos, el tratamiento de términos multipalabra, así como la resolución de acrónimos y abreviaturas son algunos de los siguientes pasos dentro de la planificación de nuestro trabajo.

La evaluación del sistema fue realizada por un farmacéutico, debido a la falta de corpus etiquetados para el dominio farmacológico. Este proceso manual, además de tedioso, implica una gran cantidad de tiempo y esfuerzo. Por este motivo, con el objeto de reducir la carga de nuestro experto, hemos supuesto que la información aportada por UMLS es correcta.

Sin embargo, una revisión manual de una pequeña muestra de los conceptos clasificados como sustancias farmacológicas en UMLS, mostró que algunos de ellos no eran sustancias, sino acciones o funciones farmacológicas. Esta inconsistencia semántica también fue reportada Schulze-Kremer y colegas (Schulze-Kremer et al., 2004). Por tanto, somos conscientes que es imprescindible evaluar manualmente el conjunto de conceptos clasificados por UMLS para conseguir una estimación real de la precisión y cobertura del sistema.

Integrar un modulo para el reconocimiento de entidades del dominio general, así como una lista de términos biomédicos no incluidos en UMLS son algunas de las medidas futuras para reducir el coste de la evaluación.

Agradecimientos

Los autores agradecen a María Segura Bedmar, responsable del centro de información de medicamentos del Hospital de Móstoles, su valiosa ayuda en la evaluación del sistema.

Bibliografía

Ananiadou, S. 1994. A Methodology for Automatic Term Recognition. En: *Proceedings of COLING-94*. Kyoto, Japan. 1034-1038

Chiang, J.-H. and Yu, H.-C. 2003. Meke: Discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics, Vol. 19(11): 1417–1422.

Collier N, Takeuchi K. 2004. Comparison of character-level and part of speech features for name recognition in biomedical texts:423–35.

The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* 2003;31(1):172–5.

Friedman, C., Shagina, L., Lussier, Y. and Hripcsak, G., 2004. Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. 11, 392–402

Fukuda, K., A. Tamura, T. Tsunoda, and T. Takagi. 1998. "Toward information extraction: identifying protein names from biological papers". In: *Proceedings of Pac Symp Biocomput.:* 707-718.

Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. 2003. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*;19(1):135–43.

Hobbs JR. 2002. Information extraction from biomedical text. *J Biomed Inform;35*(4):260–4.

Hirschman L, Morgan AA, Yeh AS. 2002. Rutabaga by any other name: extracting biological names. *J Biomed Inform*;35(4):247–59.

Lee KJ, Hwang YS, Kim S, Rim HC. 2004. Biomedical named entity recognition using two phase model based on SVMs. *J Biomed Inform.* 37(6):436–47.

Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. In: Proceedings of Pacific Symposium on Biocomputations. 2003. pp. 427–38.

Rindflesch, T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac. Symp. Biocomput. 5, 517–528

Smith JW, Seidl LG y Cluff LE, 1969. Studies on the epidemiology of adverse drug interactions. V. Clinical factors influencing susceptibility. *Ann Intern Med: 65, 629* .

Stockley, I. 2004. Stockley Interacciones farmacológicas. *Pharma Editores.* Barcelona.

Tanabe, L. y Wilbur, W.J. 2002. Tagging gene and protein names in biomedical text. Bioinformatics 18, 1124–1132

Tsuruoka Y, Tsujii J. 2003. Boosting precision and recall of dictionarybased protein name recognition. *En: Proceedings of NLP in Biomedicine, ACL*. Sapporo, Japan; 41–8.

Wilbur WJ, Hazard GF Jr, Divita G, Mork JG, Aronson AR, Browne AC. 1999. Analysis of biomedical text for chemical names: a comparison of three methods. Proc. AMIA Symp. 176–180

Zhang J, Shen D, Zhou G, Su J, Tan CL. 2004. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *J Biomed Inform. 37(6):*411–22.

Schulze-Kremer S, B. Smith, A. Kumar. 2004. Revising the UMLS Semantic Network. In: Fieschi M, Coiera E, Li Y-C, editors. *Proceedings of Medinfo*. San Francisco, CA; 2004. p. 1700.