ARTÍCULO

Reconocimiento y anotación de nombres de fármacos genéricos en la literatura biomédica

Recognizing and annotating generic drug names in biomedical literature

Prof. Carmen Gálvez
Departamento de Información y Comunicación de la Universidad de Granada, España.

RESUMEN

Este trabajo propone un sistema para la identificación y anotación de nombres de fármacos genéricos en textos biomédicos basado en modelos de estado -finito. El procedimiento presentado utiliza reglas de nomenclatura para fá rmacos genéricos, recomendadas por el Consejo *United States Adoptated Names* (USAN), que van a permitir la clasificación de los fármacos en familias farmacológicas, y una herramienta de ingeniería lingüística basada en tecnología de estado -finito. Por medio de una interfaz gráfica, se han construido analizadores capaces de identificar, clasificar y etiquetar nombres de fármacos genéricos, utilizando los afijos recomendados por USAN. El sistema consigue un 99,8 % de precisión y un 92 % de exhaustividad sobre una colección de 259 resúmenes de artículos científicos extraídos de la base de datos Medline. La combinación de reglas USAN y tecnología de estado-finito constituye un procedimiento eficaz para la detección, clasificación y etiquetado de nombres de fármac os genéricos.

Palabras clave: fármacos genéricos, nombres de fármacos, reconocimiento de entidades biomédicas, interacciones farmacológicas.

ABSTRACT

This paper proposes a system for identification and annotation of generic drug names in biomedical te xts based on finite -state models. The proposed procedure uses naming rules for generic drugs, recommended by the *United States Adoptated Names* (USAN) Council, allow the classification of drugs in drug families, and linguistic engine based on finite -state techniques. Through a graphical interface, we have built analyzers able to identify to identify, classify and assign annotations to generic drug names, using affixes recommended by USAN. The evaluation corpus consists of 256 Medline abstracts. The system ac hieves a 99.8% precision and 92% recall. The combination of rules USAN and finite -state technology is an effective procedure for the detection, classification and tagging of generic drug names.

Key words: generic drugs, drug naming, biomedical named entit ies, drug-drug interactions.

INTRODUCCIÓN

El reconocimiento y clasificación de nombres de fármacos constituye la primera etapa en el desarrollo de los sistemas dirigidos a la extracción automática de interacciones farmacológicas de la literatura biomé dica. Los últimos avances en biomedicina han provocado un incremento vertiginoso del número de publicaciones científicas. Por este motivo, es necesario el desarrollo de sistemas que faciliten la extracción de conocimiento y un acceso eficiente a la informa ción en el dominio de la biomedicina. La identificación, clasificación y anotación de las entidades biomédicas es el primer paso en progreso de tales sistemas. En este sentido, la identificación de nombres de fármacos genéricos es una tarea compleja, tenie ndo en cuenta los problemas que implica el procesamiento del texto farmacológico.

Las interacciones farmacológicas están provocadas por modificaciones o alteraciones cuantitativas o cualitativas del efecto de un fármaco, causadas por la administración sim ultánea o sucesiva de otro fármaco, pero también a través de plantas medicinales, alimentos, bebidas o contaminantes ambientales. [1] Esta modificación suele traducirse en una variación de la intensidad, aumento o disminución del efecto habitual o en la apari ción de un efecto distinto al esperado. El riesgo de aparición de una interacción farmacológica aumenta en función del número de fármacos administrados al mismo tiempo a un paciente. Si un paciente que toma dos fármacos ve aumentado el efecto de uno de ell os se puede caer en una situación de sobredosis y, por tanto, de mayor riesgo de que aparezcan efectos secundarios. A la inversa, si ve su acción disminuida se puede encontrar ante falta de utilidad terapéutica por infradosificación. Por ejemplo, los pacie ntes que reciben paracetamol pueden aumentar su efecto analgésico si se les administra codeína. Del mismo modo, la combinación de ácido clavulánico y la amoxicilina puede evitar la resistencia de las bacterias al antibiótico. Así, la polimedicación facilit a la aparición de interacciones cuyo resultado puede ser una reacción adversa o la pérdida de efecto terapéutico; de aquí se deduce la importancia de la identificación de interacciones en el campo de la medicina. [2-3]

Aunque la incidencia de la interacción farmacológica es difícil de determinar, se relaciona fundamentalmente con el número de fármacos administrados conjuntamente al mismo paciente. Conocer las interacciones de interés clínico y sus mecanismos de producción podría ayudar a identificarlas y prev enirlas. Por otra parte, la investigación y el desarrollo de medicamentos requiere esfuerzos coordinados de múltiples disciplinas; entre ellas, las experimentales, como la genómica, la proteómica, la química farmacéutica o la farmacología, se unen con las distintas especialidades médicas y con la epidemiología. Todas estas disciplinas se conectan en proyectos de gran complejidad que permiten avanzar desde el descubrimiento de nuevas dianas terapéuticas hasta la puesta en el mercado de nuevos medicamentos.

Los profesionales de la salud utilizan distintas bases de datos para identificar interacciones farmacológicas, que incluyen además información sobre el mecanismo de acción y gravedad de las posibles reacciones adversas, tales como:

- Base de datos *Micromedex,*[4] que contiene además información relevante respecto a interacciones farmacológicas: a) *DRUG-REAX*: recoge las interacciones medicamentosas y reacciones adversas notificadas en la literatura biomédica o por los titulares de comercialización del medicament o, y b) *DRUGDEX*: recoge monografías de medicamentos con amplia información basada en la evidencia disponible en cuanto a utilidad terapéutica y comparación con principios activos similares, incluyendo además información de interés sobre interacciones.

- Base de datos *Lexi-Comp,*[5] que mediante un sistema de acceso sencillo es capaz de responder a las preguntas más comunes, como la verificación de la dosis, advertencias, precauciones o reacciones adversas, así como una visión del contenido referencial, para ayu dar a los farmacéuticos clínicos en la formulación de recomendaciones de tratamientos.

En diversos trabajos realizados, se ha demostrado que la calidad de las bases de datos de interacciones es muy desigual, y en consecuencia es muy difícil concretar la relevancia clínica de cada una de las interacciones. [6-8] Frente a esta situación, los textos biomédicos constituyen el principal recurso para obtener información sobre nuevas interacciones.

Por tanto, la literatura biomédica constituye la fuente de informac ión científica por excelencia, así como los resúmenes de la base de datos *Medline*, producida por la *National Library of Medicine* (NLM) de Estados Unidos. Nunca antes los médicos han tenido un acceso tan fácil e inmediato al último hallazgo científico como ahora, pero tampoco nunca antes el exceso de información se ha convertido en una barrera para que los profesionales puedan tomar la mejor decisión, sabiendo que lo hacen a partir de las mejores evidencias disponibles. La información médica es cada vez mayor, y la base de datos *Medline* se ha convertido en la fuente de información biomédica más utilizada; sin embargo, a pesar de la accesibilidad a estos recursos, la extracción automatizada de información útil sigue suponiendo un desafío, ya que los textos y resúmenes están en lenguaje natural. Por tanto, el uso de recursos y tecnologías de procesamiento del lenguaje podría facilitar el acceso a la información en el dominio farmacológico.

En relación con lo anterior, la extracción de información (EI) tiene un papel fundamental, como disciplina perteneciente al procesamiento del lenguaje natural (PLN), ya que utiliza conjunto de técnicas para la obtención de datos estructurados y no-ambiguos del lenguaje natural con diferentes propósitos, tales como la construcción de bases de datos, o aplicaciones relacionadas con la recuperación de información (RI).[9] La EI es esencial para analizar y extraer información útil de los textos biomédicos, imposible de realizar de forma manual. Por esta última razón, son muchos los trabajos dedicados a la investigación sobre el empleo de técnicas de EI a los textos biomédicos. [10-12]

Otros métodos de procesamiento y acceso a la información, los constituyen las técnicas de minería de datos, *data-mining*, y la minería de texto o minería t extual, *text-mining*. Estas tecnologías surgen como métodos emergentes que sirven de soporte para el descubrimiento de conocimiento que poseen los datos almacenados. La minería de datos se define como el descubrimiento de conocimiento, a partir de patrones observables de datos estructurados en bases de datos relacionales. Se le denomina comúnmente *Knowledge-Discovery in Databases* (KDD). La minería textual está orientada a la extracción de conocimiento a partir de datos no - estructurados en lenguaje natural al macenados en las bases de datos textuales. Se identifica con el descubrimiento de conocimiento en los textos y se le denomina comúnmente *Knowledge-Discovery in Text* (KDT). Tanto la minería de datos como

la minería de texto son técnicas de análisis de información. En el caso de la información biomédica, mediante el proceso de análisis se le agrega valor a la información hasta convertirla en conocimiento. Solo las computadoras pueden manipular rápidamente la gran cantidad de datos.[13] La minería de texto es una herramienta de análisis encargada del descubrimiento de conocimiento que no existía explícitamente en ningún texto de la colección, pero que surge al relacionar el contenido de varios de ellos.[14] La minería de texto adopta un enfoque semiautomático, y establece un equilibrio entre el análisis humano y el automático (antes de la etapa de descubrimiento de conocimiento es necesario procesar de forma automática la información disponible en grandes colecciones documentales y transformarla en un formato que facilite su comprensión y análisis).

Desde el punto de vista tecnológico, el procesamiento de grandes volúmenes de información biomédica en texto libre no-estructurado requiere la aplicación de una serie de técnicas de análisis, tales como la identificación, extracción y anotación de entidades biomédicas, así como el descubrimiento de conocimiento y visualización de datos. En este contexto, las técnicas basadas en PLN permiten mejorar la utilización de la lengua en los sistemas informáticos, asimilando, analizando, seleccionando y presentando la información para contribuir a superar el problema de exceso de información.

Por constituir la identificación de los nombres de fármacos una tarea esencial en los sistemas de EI útil en farmacología, el objetivo se este trabajo es proponer un procedimiento basado en una combinación de recursos y técnicas de PLN para la identificación de nombres de fármacos genéricos en la literatura biomédica. Nuestra propuesta se basa en la aplicación de:

- Reglas de nomenclatura para fármacos genéricos, recomendadas por el Consejo *United States Adoptated Names* (USAN) que van a permitir la clasificación de los fármacos en familias farmacológicas.

- Una herramienta lingüística que utiliza una amplia gama de dispositivos de cómputo que utilizan tecnología de estado-finito.

REGLAS DE NOMENCLATURA PARA FÁRMACOS GENÉRICOS

Los fármacos son sustancias químicas que presentan una acción biológica, lo cual no significa que puedan ser siempre utilizadas con fines terapéuticos. Mientras que los medicamentos son sustancias químicas que se utilizan con fines terapéuticos, es decir, todos los medicamentos, son fármacos; pero no todos los fármacos son medicamentos. Además, un medicamento puede estar constituido por uno o varios fármacos. Un fármaco tiene tres nombres:

- *Nombre químico*, que se refiere a la composición molecular del fármaco y debe seguir las reglas de la nomenclatura química.

- *Nombre genérico* o nombre oficial del fármaco durante su existencia, establecido por organismos oficiales nacionales e internacionales. Se trata de un nombre de titularidad pública y que no está protegido por patente.

- *Nombre comercial* o marca, que es el nombre dado por la compañía farmacéutica que lo comercializa. Se trata del nombre registrado o de la patente y consiste en la protección que se da oficialmente para explotar de modo industrial un fármaco.

Para la denominación oficial de los fármacos genéricos, contamos con la *Denominación Común Internacional* (DCI) de los principios activos, establecida por la Organización Mundial de la Salud (OMS) a nivel internacional. Cada DCI es un nombre único que es reconocido a nivel mundial y es de propiedad pública. Son nombres independientes de los laboratorios y no tienen propietario, de manera que pueden ser usados sin restricción alguna. Se recomienda que sean nombres muy simples, debido a la generalización de su uso internacional. Las DCI deben tender a mantener un parentesco con otras sustancias que pertenezcan al mismo grupo farmacológico. Sin embargo, la implantación de las DCI no es universal. Existen organismos que regulan los nombres a nivel nacional (tabla 1). Estos organismos adoptan los nombres de las DCI y los adaptan a la lengua de cada país.

Tabla 1. Organismos nacionales reguladores de nombres de fármacos genéricos

País	Denominación común nacional	Significado
Estados Unidos	USAN	*United States Accepted Name*
Gran Bretaña	BAN	*British Approved Name*
Francia	DCF	*Denomination Commune Française*
España	DOE	*Denominación Oficial Española*
Italia	DCIT	*Denominazione Comune Italiana*
Países Nórdicos	NFN	*Nordiska Farmakopknamnden*
Japón	JAN	*Japanese Accepted Name*

La OMS ha aprobado partículas, tanto prefijos como sufijos, específicas para los distintos grupos farmacológicos. Las prácticas habituales para nombrar fármacos recaen en el uso de afijos. Estos afijos permiten clasificar los fármacos en familias farmacológicas según su estructura química. La lista recomendada por el Consejo USAN representa los afijos comunes establecidos para cada parámetro químico o farmacológico. Por ejemplo, los antiinflamatorios podrían contener alguno de los siguientes afijos: *-ac*, *-bufen*, *-butazone*, *-fenamic*, *-icam*, *-metacin*, *-nidap*, *-nixin*, *-profen*, *sal-*, *-sal-* y *sal*.

Estos afijos, reglas de nomenclatura y sus definiciones aprobados por el Consejo USAN se recomiendan para que se acuñen en los nuevos nombres de fármacos que pertenezcan a una serie establecida de agentes relacionados. De este modo, se proporciona un reconocimiento inmediato de los compuestos similares pertenecientes a una misma familia farmacológica. La lista de afijos no es exhaustiva, ya que no incluye todos los afijos utilizados por el Consejo USAN ni otros grupos de nomenclaturas nacionales o internacionales. Además, hemos de tener en cuenta que constantemente nuevos afijos se pueden crear y que otros existentes se pueden modificar.

En el procedimiento presentado en este trabajo, hemos adoptado la clasificación de los fármacos, según sus afijos, recomendada por USAN. [15] En la tabla 2 se muestra algunos de estos afijos.

Tabla 2. Algunos afijos recomendados por el Consejo *United States Adopted Names* (USAN)

Afijos	Definición	Ejemplos
-ac	anti-inflammatory agents (acetic acid derivatives)	bromfenac
-adol	analgesics (mixed opiate receptor agonists/antagonists)	levonantradol
-adox	antibacterials (quinoline dioxide derivatives)	carbadox
-aril-	antiviral (arildone derivatives)	fosarilate
-azosin	antihypertensives (prazosin type)	doxazosin
-bactam	beta-lactamase inhibitors	sulbactam
-butazone	anti-inflammatory analgesics (phenylbutazone type)	mofebutazone
-carbef	antibiotics (carbacephem derivatives)	loracarbef
-cillin	penicillins	ampicillin
-ectin	antiparasitics (ivermectin type)	doramectin
-mulin	antibacterials, pleuromulin derivatives	retapamulin
-oxacin	antibacterials (quinolone derivatives)	difloxacin
-oxanide	antiparasitics (salicylanilide derivatives)	bromoxanide
-oxef	antibiotics (oxacefalosporanic acid derivatives)	flomoxef
-prim	antibacterials (trimethoprim type)	ormetoprim
-profen	anti-inflammatory/analgesic agents (ibuprofen type)	flurbiprofen
-quiline	antibiotics, diarylquinoline structure	bedaquiline
-virsen	antivirals	afovirsen
-tide	peptides	octreotide

TÉCNICAS BÁSICAS DE PROCESAMIENTO DEL LENGUAJE NATURAL

A través de tecnologías basadas en PLN se construyen herramientas automáticas con suficiente información lingüística en forma de reglas y patrones que permite realizar numerosas actividades. Los sistemas de P LN deben identificar todos los

niveles de la lengua: nivel morfológico, léxico, sintáctico. La mayoría de técnicas de procesamiento del lenguaje se desarrolla por medio de diferentes etapas que pueden operar de manera secuencial o paralela, tales como: a) pre-procesamiento textual; b) análisis morfológico; y c) análisis sintáctico o *parsing*.[16]

La primera etapa de pre-procesamiento de cualquier sistema de PLN tiene lugar en el nivel textual. En este nivel, el texto puede ser considerado como una simple secuencia de caracteres. Las tareas básicas que deben abordarse a este nivel son: la segmentación del texto, y la localización de unidades léxicas o palabras. Localizar las palabras ortográficas constituye una tarea sencilla si el espacio o los signos de puntuación actúan como separadores.

El siguiente paso en el tratamiento de la lengua consiste en el análisis morfológico. Esta tarea es normalmente realizada por un analizador morfológico cuyo papel es el de recuperar la morfología de las palabras; es decir, las formas con que se construyen las palabras a partir de unidades significativas más pequeñas, llamadas `morfemas'. Los morfemas se clasifican en dos clases: morfema raíz o lema (*stem*) y afijos. Generalmente, las palabras se forman a través de mecanismos de flexión, derivación o composición a partir de sus formas canónicas. La tarea de descomposición de una palabra de la entrada en su forma de base y sus afijos se denomina *stemming* o lematización.

Un analizador morfológico debe constar por lo menos tres partes: un diccionario o lexicón con la lista de los lemas; una lista de afijos con sus reglas de orden, ya que los afijos no pueden aparecer en un orden arbitrario, y un conjunto de reglas ortográficas en el caso de que la adición de un afijo las requiera. Para que el procesamiento morfológico sea posible, cada lema debe ser previamente etiquetado. Se denomina `etiquetado', *POS tagging* (*part-of-speech tagging)* al procedimiento de asignar a cada una de las unidades léxicas presentes el conjunto de sus categorías gramaticales posibles.[16] El objetivo de un etiquetador es el de asignar a cada palabra la categoría más `apropiada' dentro de un contexto. Existen tres grandes procedimiento de etiquetado:

- *Técnicas de etiquetado basadas en reglas*. Los etiquetadores basados en reglas utilizan conocimiento lingüístico, generalmente expresado en forma de reglas o restricciones para establecer las combinaciones de etiquetas aceptables o prohibidas. Las reglas se escriben manualmente, responden a criterios lingüísticos y se representan en forma explícita. Otros métodos se enfrentan al problema de la variabilidad del lenguaje desde una aproximación lingüística, por medio de técnicas cuyo objetivo es la reducción de las variantes léxicas a lemas. En esta línea, una de las implementaciones computaciones más importantes la constituyen los analizadores basados en tecnología de estado-finito:[17,18]

- *Técnicas de etiquetado basadas en métodos estadísticos o probabilísticos*. Estos etiquetadores se basan en la evidencia empírica obtenida de corpus lingüísticos voluminosos. El problema de estos sistemas reside en el aprendizaje del modelo estadístico utilizado. Se han utilizado técnicas de aprendizaje supervisado partiendo de *corpus* etiquetados manualmente y técnicas de aprendizaje no supervisado en las que no es precisa esa intervención manual. Un algoritmo clásico utilizado para el etiquetado estadístico es el de los Modelos Ocultos de Markov (*Hidden Markov Models*). Este enfoque se caracteriza por asumir que la probabilidad de una cadena de símbolos puede ser calculada en base a sus partes o *n-gramas.* El modelo de *n-gramas* más básico es el de los *unigramas*; es decir, la búsqueda de la etiqueta más probable para cada palabra o *token.* Para esto, es necesario entrenar el sistema con un *corpus* etiquetado previamente.[19]

- *Técnicas de etiquetado híbridas*, que combinan tanto los métodos basados en reglas como los estadísticos para intentar recoger los aspectos positivos de cada una de ellas y evitar sus limitaciones. Un sistema de este tipo fue introducido por *Brill*[20] y se basa en el aprendizaje automático. Cada palabra se rotula con la etiqueta más probable, luego se cambia la etiqueta aplicando reglas del tipo `*si la palabra -1 es un determinante cambie la etiqueta a nombre*' y se reetiqueta la palabra. Se obtiene de esta manera una secuencia de reglas de transformación de etiquetas.

Una vez analizado y etiquetado tal texto de forma total o parcial, puede realizarse el análisis sintáctico (*parsing*). Se trata de un proceso por medio del cual se convierte el texto de entrada en otras estructuras, comúnmente denominadas `*árboles*', que son más útiles para el posterior análisis y capturan la jerarquía implícita de la entrada. Durante el procesamiento se producen distintas estructuras intermedias o de trabajo, hasta producir un árbol de análisis estructural de la secuencia de entrada.[21] Hay diferentes técnicas y algoritmos de *parsing.* Estas se pueden agrupar básicamente entre tres tipos diferentes:

- *Procesamiento paralelo o secuencial*. Se refiere fundamentalmente a dos tipos de análisis de secuencias. La técnica de procesamiento en paralelo prueba diferentes posibilidades de combinación en paralelo y guarda la pista de los estados posibles. Frente a este, la estrategia de procesamiento secuencial prueba primero una posibilidad hasta el final, y si no tiene éxito, retrocede al punto de partida y prueba otra ruta hasta dar con la estructura que corresponde a la secuencia de la entrada.

- *Procesamiento descendente o ascendente*. Se refiere al punto de partida del árbol estructural que el *parser* debe construir. Si se está procesando una oración, en la parte superior se representa a la oración en su totalidad y, en la parte inferior del árbol hay nodos que representan los elementos léxicos individuales o palabras. La dirección ascendente y la descendente dependen del punto de partida: si comienza el procesamiento en la parte superior de la oración y va dividiendo la entrada progresivamente en partes cada vez más pequeñas, hasta llegar a las palabras, será un *parser* descendente (*top-down-parser).* El *parser* será ascendente *(bottom-up)* si, por el contrario, el análisis comienza por los elementos léxicos individuales y culmina con la oración en su totalidad.

- *Procesamiento determinista/no-determinista*. Se refiere al carácter guiado o no guiado del modelo. Es decir, si el modelo no permite decidir qué regla de la gramática se aplicará en un momento determinado, se tratará de un modelo no-determinista; en cambio, si se utilizan mecanismos que conducen a un resultado concreto sin vacilaciones, se hablará de un procesamiento determinista.

Sin embargo, las técnicas de análisis sintáctico tienen dos grandes problemas propios de los analizadores automáticos: la ambigüedad y el costo informático que implica el tiempo de procesamiento, que suele ser muy lento y costoso. Para solucionar estos problemas se puede realizar un análisis superficial o fragmental (*shallow parsing*) en lugar de un análisis en profundidad. Para muchas aplicaciones no es necesario desarrollar un análisis del texto completo.

El objetivo de los analizadores fragmentales, también denominados agrupadores sintácticos o *chunkers*, es la detección de determinados segmentos textuales, tales como de frases nominales, determinados nombres o entidades. En estos casos, es

frecuente el uso de técnicas de estado-finito y la actuación de transductores en cascada.[22]

En este trabajo se van a utilizar modelos de estado-finito tanto para el análisis morfológico como sintáctico de las entidades biomédicas; en este caso, nombres de fármacos genéricos.

APROXIMACIÓN A LOS MODELOS DE ESTADO-FINITO

La teoría de los lenguajes formales se dirige a aquellas expresiones que pueden ser descritas de forma muy precisa, como son los lenguajes de programación. Los lenguajes naturales no son lenguajes formales, y, por tanto, no hay un límite claramente definido entre una sentencia correcta de otra que no lo es. Sin embargo, se pueden adoptar algunas aproximaciones formales a ciertos fenómenos del lenguaje natural susceptibles de una codificación similar a la realizada en los lenguajes de programación. Estas descripciones formales se utilizan por los lingüistas computacionales para expresar teorías sobre aspectos específicos de los lenguajes naturales, tales como el análisis morfológico y el análisis y etiquetado de segmentos de texto.

Johnson[23] fue el primero en observar que determinadas morfológicas se podrían representar por mecanismos de estado-finito, denominando a su formalismo `*two level model*'. La idea del modelo de dos-niveles fue clave para el progreso del formalismo computacional sobre la morfología propuesto por *Koskenniemi*.[24] El modelo de *Koskenniemi* estableció una correspondencia entre la forma canónica, o forma léxica, y la forma superficial de las palabras. Esta relación la representó usando transductores finitos.

De forma sintetizada, un transductor de estado-finito (FST, siglas en inglés), es un sistema de representación computacional que comprende un conjunto de estados y una función de transición, que define el cambio de estado. La función de transición se etiqueta con un par de símbolos que constituyen el alfabeto del *input* y el alfabeto de *output*. Este mecanismo se puede representar en la forma de un diagrama o gráfico de estado-finito. El transductor tomaría cadenas en el *input* y las relacionaría con cadenas en el *output*. Formalmente un FST se define como una tupla de cinco elementos que se expresa de la forma siguiente: [25]

$FST = (\Sigma, Q, i, F, E)$

donde:

Σ = alfabeto de input y output
Q = número de estados
i = estado inicial
F = estado final
E = número de relaciones de transición

En la figura 1 se muestra la representación gráfica de un transductor cuyos arcos están etiquetados con pares de símbolos que constituyen el alfabeto de *input* y *output*. Por ejemplo, "*a*" denota el símbolo superior y "*b*" el símbolo inferior.

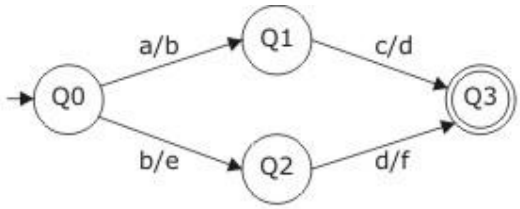

Fig. 1. Representación gráfica de un transductor de estado-finito.

Este transductor podría establecer una relación entre el lenguaje superior y el inferior. Así, este mecanismo podría reconocer la cadena representada por " *ac*" y la podría transformar en la cadena "*bd*". La equiparación es bidireccional, y una cadena de un lenguaje se podría corresponder a una, o más cadenas de otro lenguaje. Las transducciones son posibles si la cadena en la parte del *input* lleva al transductor a un estado final.

La aplicación del formalismo de estado-finito a la unificación y etiquetado de términos parte básicamente de que se puede establecer una relación de equivalencia entre las distintas formas superficiales y la raíz, o lema, a la que se le puede añadir una etiqueta de la categoría gramatical correspondiente, o etiqueta POS (*part-of-speech).* Esta correspondencia se puede implementar computacionalmente por medio de transductores.[26] Una analizador de *dos-niveles* o *lematizador* desarrollado con tecnología de estado-finito se encargaría de equiparar formas variantes léxicas, a formas unificadas, tal y como se representa en la figura 2.

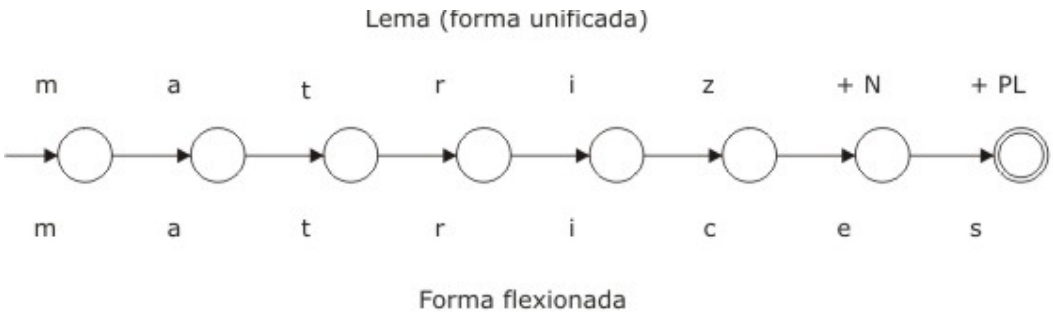

Fig. 2. Relación entre la raíz de un término y sus variantes representadas en un transductor de estado-finito.

De la misma forma, se ha utilizado el formalismo de estado-finito para el análisis superficial (*shallow parsing*), donde lo que se intenta es recuperar solo una parte de la información sintáctica del texto. Estas aproximaciones se han basado principalmente en realizar el análisis y etiquetado de segmentos de texto a través de cascadas de transductores, donde cada transductor agrega información sintáctica dependiendo del contexto.[27] El etiquetado gramatical consiste en asociar a cada palabra la categoría gramatical a la que pertenece. Esta tarea suele ser una de las primeras etapas en cualquier sistema de procesamiento de textos. La mayor dificultad de este problema viene provocada por la ambigüedad que presentan numerosas palabras, que pueden tener diferentes funciones gramaticales. Esta ambigüedad hace que la solución al etiquetado gramatical sea compleja y que pase por el uso de la información que proporciona el contexto de cada palabra.

El análisis y etiquetado del texto se realiza a través de cascadas de transductores, donde cada transductor agrega, o modifica, información previamente generada por

los transductores de la cascada. Esta técnica se desarrollaría básicamente en cuatro fases:

- Cargar el texto original, representado como una secuencia de *tokens*.
- Reconocimiento y etiquetado de las raíces, y terminaciones, de las palabras en un grafo de texto.
- Realización de sucesivas pasadas sobre el grafo de texto, aplicando en cada una de ellas módulos de reglas. La aplicación de las reglas recorre el grafo de texto de izquierda a derecha, analizando cada una de las posiciones si el *ítem* coincide con la categoría, que toda regla según la implementación debe especificar.
- Generación de la salida a partir de grafo etiquetado.

HERRAMIENTA LINGÜÍSTICA BASADA EN MODELOS DE ESTA DO-FINITO

Siguiendo con el planteamiento anterior, en nuestra propuesta hemos utilizado un software de ingeniería lingüística, basado en modelos de estado -finito,[28] denominada NooJ.[29] Se trata de una herramienta de libre acceso capaz de formalizar e identificar distintas unidades lingüísticas de forma automáticas, tales como análisis morfológico y etiquetado de palabras, análisis sintáctico y reconocimiento de entidades. Las descripciones de las lenguas naturales se formalizan en diccionarios electrónicos y gramáticas representadas por conjuntos de gráficos. Este recurso permite, además, aplicar sofisticadas consultas lingüísticas a los textos con el objetivo de crear índices y concordancias, anotar y etiquetar automáticamente textos o realizar análisis est adísticos. NooJ incluye herramientas para construir, depurar, mantener y acumular grandes conjuntos de recursos lingüísticos, y se puede aplicar a los textos de gran tamaño. Una de las características de NooJ es que puede procesar varios tipos de unidades lingüísticas en los textos. Esta herramienta lingüística utiliza un sistema de etiquetado que se puede aplicar, en todos los niveles de análisis, permitiendo la formalización de diversos fenómenos lingüísticos de forma independiente. Todas las unidades lingüísticas reconocidas por los analizadores léxicos, sintácticos y semánticos de NooJ se representan en forma de anotaciones.30

Una etiqueta o anotación es un par *posición-información* que indica que cierta secuencia del texto tiene determinadas propiedades . La Interfaz Gráfica de NooJ (*Graphical User Interface*) nos ofrece la posibilidad de construir los analizadores de estado-finito léxicos y sintácticos, que operan en cascada en todos los niveles de la formalización. De forma sintética, la herramienta trab aja con la siguiente estrategia de procedimiento para conseguir la identificación y etiquetado de los nombres de fármacos genéricos:

- *Preprocesamiento del texto,* que ejecuta la segmentación del texto en palabras, dígitos y delimitadores textuales.
- *Análisis morfológico de estado-finito*, que tiene como *input* las unidades lingüísticas (*Atomic Linguistic Units, ALU*), tales como raíces y afijos de los nombres de fármacos genéricos, y producir, como *output,* un conjunto de etiquetas, representadas en forma de ano taciones en el texto (Text Annotation Structure, TAS). Estas anotaciones están siempre sincronizadas con el fichero de texto original, que nunca se modifica.

- *Análisis sintáctico de estado-finito*, que tiene como *input* las unidades lingüísticas anotadas, identificadas en el proceso anterior y como *output* la correspondiente etiqueta.

Esta arquitectura requiere que los analizadores de NooJ se comuniquen a través de diversas anotaciones en el texto; es decir, que cada estructura anotada se almacena por el sistema y los resultados pasan al analizador siguiente. Por otra parte, esas anotaciones, en los textos, se pueden etiquetar y exportar finalmente como un documento XML.

MÉTODOS

El primer paso para desarrollar el procedimiento presentado en este trabajo es obtener una muestra de textos biomédicos, que actúen de *corpus* para la extracción de los nombres de fármacos genéricos. Aunque la literatura biomédica es la fuente de información científica por excelencia, nos restringimos a resúmenes o *abstracts* de la base de datos *Medline*. Se ha trabajado con una colección de 259 resúmenes de artículos científicos de *Medline* recuperada mediante una búsqueda de nombres de familias farmacológicas definidas como analgésicos. Tomando los afijos recomendados por USAN, realizamos la siguiente *query*: "**adol OR *butazone OR *fenine OR *eridine OR *fentanil*" con el límite adicional del siguiente período: "desde el 1ro. de enero de 2011 hasta el 31 de diciembre de 2011".

A partir del material anterior, el método propuesto consta de cuatro módulos que se ejecutan en cascada utilizando, como ya se ha mencionado, la herramienta lingüística NooJ:

1. Procesamiento de los resúmenes extraídos de la base de datos *Medline.*

2. Construcción de la gramática morfológica en forma de transductor , utilizando la Intefaz Gráfica de NooJ, que nos va a permitir la identificación de las unidades lingüísticas más pequeñas, ALU (*Atomic Linguistic Units*), es decir, raíces y afijos contenidos en los nombres de fármacos genéricos, así como su representación en forma de anotaciones textuales, o TAS (*Text Annotation Structure*).

3. Construcción del analizador sintáctico en forma de transductor, utilizando la Intefaz Gráfica de NooJ, que nos facilitará la identificación de los nombres de fármacos genéricos, a los que se les asignó una etiqueta.

4. Exportación de los resúmenes con los nombres de los fármacos genéricos anotados y clasificados.

En primer lugar, importamos el fichero con los resúmenes a la herramienta NooJ. En esta fase, los resúmenes se dividen en oraciones, se identifican los tokens, palabras, dígitos y delimitadores.

El segundo paso es el diseño y construcción del analizador léxico capaz de reconocer las unidades lingüísticas, ALU, definidas aquí como afijos, que clasifican los fármacos en familias farmacológicas. La gramática morfológica que identifica los nombres de fármacos y sus afijos se representa por medio de gráficos. NooJ proporciona herramientas para reconocer y describir dichas unidades lingüísticas mediante gráficos, que internamente se compilan en transductores de estado-finito. Un grafo es un conjunto de nodos conectados, en el que se distingue un nodo inicial

y un nodo final. Con el fin de describir y representar estas secuencias se sigue una ruta o camino, es decir, una secuencia de conexiones que comienza en el nodo inicial del gráfico, y termina en el nodo final.

Utilizando la Interfaz Gráfica de NooJ, hemos diseñado un gráfico de estado -finito (Fig. 3) en el que el símbolo <L> identifica cualquier secuencia de letras dentro de una palabra y el nodo siguiente reconoce los afijos de la familia farmacológica de los analgésicos: *-adol, -butazone, -eridine, -fenine, -fentanil* y *-profen.* Por ejemplo, el gráfico reconoce la raíz de la palabra "*pip*" (a través del símbolo <L>) y el nodo siguiente se encargaría de identificar el afijo relacionado con dicha raíz, en este caso "*eridine*"; a continuación la variable $R almacena la raíz y la asocia c on la información "DRUG+ERIDINE+analgesics".

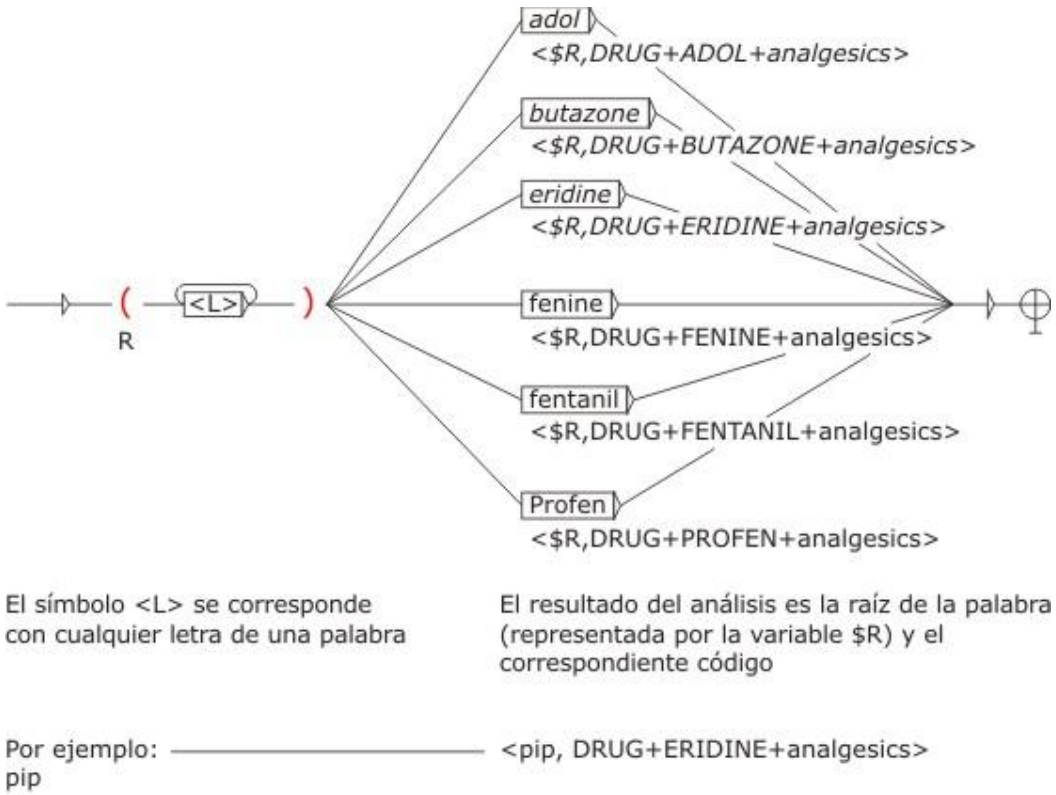

Fig. 3. Analizador léxico que procesa los afijos de los analgésicos.

A continuación, aplicamos el analizador léxico diseñado a los resúmenes del *corpus*. Como se muestra en la tabla 3, el análisis daría como resultado la identificación de los distintos afijos de los nombres pertenecientes a una familia farmacológica y la correspondiente equiparación con la anotación, o TAS (*Text Annotation Structure*).

Tabla 3. Ejemplo de la aplicación del analizador léxico a una muestra de texto

Input Text	"...Remifentanil and glucose suppress inflammation in a rat model of surgical stress".
Match in Text	"...Remi,DRUG+fentanil and glucose suppress inflammation in a rat model of surgical stress".

En tercer lugar, procedemos al diseño de un analizador sintá ctico capaz de asignar una etiqueta a los nombres de fármacos. Para esto, primero localizamos en el texto procesado la anotación <DRUG> para obtener todos los nombres de fármacos reconocidos. El resultado nos da un total de 89 nombres de fármacos genéricos diferentes reconocidos.

Nuestro propósito ahora es etiquetar los nombres de fármacos reconocidos con la categoría "DRUG". Para esto, diseñamos un analizador sintáctico por medio de un gráfico de estado-finito (Fig. 4), en el que en el nodo inicial insertamos la etiqueta "<DRUG", y en el nodo final la etiqueta ">", es decir, cada anotación identificada comenzaría donde se encuentra el carácter "<" y terminaría d onde se encuentra el carácter ">"). Además, incorporaremos información sobre la familia farmacológica a la que pertenece, añadiendo a la etiqueta la propiedad "+Analgesics".

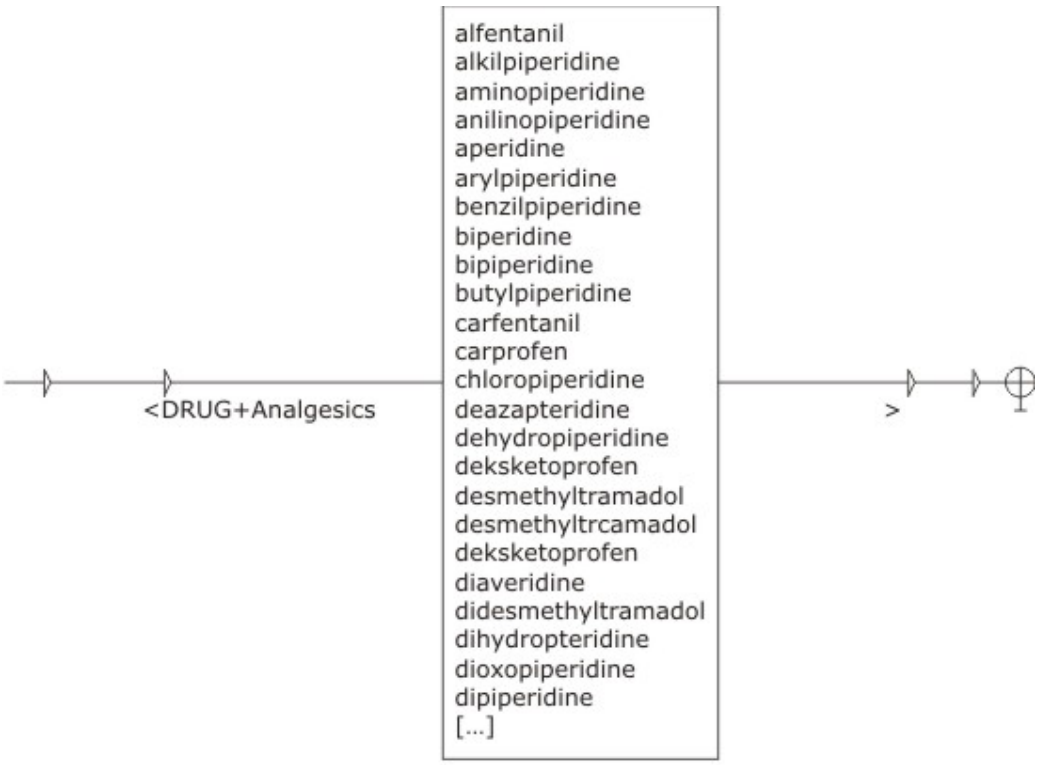

Fig. 4. Analizador sintáctico que etiqueta nombres de fármacos.

Seguidamente aplicamos el analizador a los resúmenes, en el que los nombres de pertenecientes a una familia farmacológica ya han sido reconocidos previamente. Como se muestra en la tabla 4, el análisis daría como resultado la asociación a los nombres de fármacos identificados con la etiqueta correspondiente.

Tabla 4. Ejemplo de la aplicación del analizador sintáctico a una muestra de texto

Input Text	"...Remifentanil and glucose suppress inflammation in a rat model of surgical stress".
Match in Text	"...Remifentanil<DRUG+Analgesics> and glucose suppress inflammation in a rat model of surgical stress".

Por último, una vez procesados los resúmenes con las correspondientes anotaciones asignadas a los nombres de fármacos genéricos se exportan como documentos de texto, o como un fichero XML, para su potencial aplicació n a un sistema automático de extracción de información en el dominio de la farmacología.

EVALUACIÓN Y RESULTADOS

La evaluación de los sistemas de detección y etiquetado de nombres de fármacos, que se ha presentado en este trabajo, cuenta con dos dificul tades añadidas. Por un lado, la ausencia de planteamientos similares con los que comparar la eficacia del sistema propuesto. La mayoría de los sistemas de reconocimiento de entidades biomédicas se han centrado principalmente en los nombres de los genes y proteínas. No obstante, también se han realizado trabajos sobre la detección de otro tipo de entidades como sustancias químicas31 y fármacos. [32,33] Estos últimos trabajos, aunque utilizan herramientas del PLN, se basan fundamentalmente en métodos que equiparan de forma automática los nombres de fármacos a conceptos dentro de un sistema de codificación normalizado, como Metatesauro UMLS ® (*Unified Medical Language System* ®). Por otra parte, otro gran obstáculo en la evaluación del modelo de etiquetado propuesto reside en la falta de *corpus* de evaluación. A pesar de que durante los últimos años se han desarrollado varios *corpus* biomédicos para evaluar el rendimiento de los sistemas que utilizan PLN, tales como *TREC Genomics Track,* [34,35] GENETAG,[36] BioCreative[37](*Critical Assessment of Information Extraction systems in Biology*), no disponemos de *corpus* etiquetados en el dominio farmacológico.

Teniendo en cuenta las limitaciones anteriores, la evaluación de nuestro sistema se realiza sobre una colección de textos extr aídos de la base de datos *Medline*, que está compuesto por 259 resúmenes de artículos científicos. Por otra parte, hemos utilizado los parámetros de precisión y exhaustividad (*recall)*, que son los que se emplean habitualmente en las herramientas basadas en PLN. El parámetro de precisión se define aquí como la proporción de nombres de fármacos genéricos identificados correctamente. La exhaustividad se define como la proporción de nombres de fármacos genéricos que el sistema es capaz de identificar y anotar. Incorporando estas dos métricas de evaluación, nuestro propósito es medir el grado de corrección y eficacia con el que el sistema es capaz de reconocer y etiquetar los nombres de fármacos genéricos en la literatura biomédica. Las dos medidas se calculan con las siguientes ecuaciones:

$$\text{Precisión (P)} = \frac{\text{Número de nombres de fármacos identificados correctamente}}{\text{Número total de nombres de fármacos identificados}}$$

$$\text{Recall } (R) = \frac{\text{Número de nombres de fármacos identificados correctamente}}{\text{Número total de nombres de fármacos posibles}}$$

Además, vamos a evaluar el sistema con la medida *F* (*F-Measure*) que combina en un solo valor la exhaustividad y la precisión. Se trata de una media ponderada y armónica que sirve para corregir el error de distancia en los casos en los que la exhaustividad y la precisión se compensan, de tal forma que a mayor valor de *F-Mesaure* mejor resultado. Su ecuación es:

$$F = 2\frac{P * R}{P + R}$$

Para poder aplicar los parámetros anteriores, necesitaríamos adquirir los siguientes datos:

- *Número de nombres de fármacos identificados y anotados correctamente* . Para adquirir estos datos, contrastamos cada uno de lo nombres de fármacos genéricos reconocidos por el método propuesto con la información que nos proporciona el portal de nombres de fármacos, *Drug Information Portal*, producido por *U.S. National Library of Medicine* (NLP). Además, los nombres de fármacos reconocidos se han contrastado con la información que aporta la base de datos de libre acceso *ChemSynthesis.*

- *Número total de nombres de fármacos genéricos identificados y anotados* . Para obtener estos datos aplicamos los analizadores léxicos y sintácticos, que se han diseñado, al *corpus* extraído de la base de datos *Medline*.

- *Número total de nombres de fármacos posibles existentes en el corpus*. Estos datos se obtienen por un proceso manual realizado por un experto, lo que implica una gran cantidad de tiempo y de esfuerzo, por la falta de *corpus* etiquetados para el dominio farmacológico.

El número total de fármacos genéricos reconocidos, relativos a la familia farmacológica de los analgésicos después de aplicar los analizadores, es de 2 511 *matches* en total, de los cuales 89 son diferentes (tabla 5).

Los resultados globales de la evaluación se muestran en la tabla 6. El sistema consigue una precisión de P = 99,8 %, sobre la media de F = 95 %. El total de los nombres de fármacos genéricos pertenecientes a la familia farmacológica de los analgésicos que los analizadores han identificado y anotado en el *corpus* es de 2 511, de los cuales 2 507 nombres corresponden a nombres de fármacos reconocidos correctamente. La tasa de precisión se ha visto afectada fundamentalmente por errores ortográficos, tales como *" oftramadol*" (en lugar de "*of tramadol*"), " *ofmeperidine*" (en lugar de *"of meperidine*"), " *Fsufentanil*" (en lugar de "*sufentanil*"), o "*Dmethylpteridine*" (en lugar de "*Dimethylpyridine*"). A pesar de estos errores, se puede considerar que la herramienta diseñada es bastante precisa.

Tabla 6. Resultados de la evaluación del sistema

	Precisión	Recall	F-Measure
Analizadores léxicos y sintácticos	99,8 %	92 %	95 %

La tasa de exhaustividad de los analizadores diseñados es de R = 92 %, sobre la media de F = 95 %. Esto significa que si el total de los nombres de fármacos genéricos posibles existentes en el *corpus* y susceptibles de ser identificados es de 2 711, los analizadores han conseguido identificar y anotar correctamente 2 507 nombres. El número de nombres de fármacos que los analizadores no han sido capaces de identificar es de 263 en total, que corresponden a 6 nombres diferentes: "*Piperidines", "Pteridines", "Dipiperidines", "Dioxopiperidines", "Arylpiperidines" y "butanesulfinylpiperidines"*. Es necesario aclarar que los analizadores se diseñaron para reconocer el afijo *eridine*, y no *eridines*. Por tanto, no se trata de un error de *infraanálisis* de la herramienta, sino de una falta de identificación que no estaba prevista por los analizadores. En consecuencia, se puede considerar que la cobertura del sistema es bastante aceptable.

CONCLUSIONES

Las bases de datos de interacciones farmacológicas, a pesar de contener datos muy bien estructurados, aportan una información que puede ser incompleta. Por esta razón, muchos especialistas médicos se ven obligados a revisar una gran cantidad de artículos científicos sobre seguridad de medicamentos para estar al día en todo lo publicado en relación con el tema. El desarrollo de métodos automáticos que permitan recopilar, mantener e interpretar toda esta información es crucial para la detección de interacciones entre fármacos. La primera etapa de los métodos automáticos de extracción de información es el reconocimiento y clasificación de los nombres de fármacos.

En este trabajo hemos presentado un sistema capaz de identificar y anotar nombres de fármacos genéricos en resúmenes extraídos de la base de datos *Medline*. Los resultados de la evaluación sobre la eficacia del sistema propuesto nos han llevado a las siguientes conclusiones: primero, los analizadores basados en tecnología de estado-finito consiguen identificar y anotar los nombres de fármacos genéricos con una alta precisión. Segundo, los analizadores basados en tecnología de estado-finito son capaces de identificar los nombres de fármacos genéricos con una gran cobertura. Si hay nombres que no son capaces de reconocer es porque no se han tenido en cuenta todos los posibles afijos y, por tanto, no se producen errores de *infraanálisis*.

A pesar de las dificultades que supone la evaluación del procedimiento propuesto, por la ausencia de sistemas similares con los que establecer una comparación o la falta de *corpus* etiquetados para el dominio farmacológico, los resultados preliminares muestran que se han detectado y anotado los nombres de fármacos genéricos de forma eficaz. Extender progresivamente el campo de aplicación, ampliando la cobertura de las anotaciones a través de la inclusión de un mayor número de afijos, así como integrar el sistema de reconocimiento de términos en

un proyecto más amplio de extracción automática de interacciones farmacológicas de la literatura biomédica, son algu nas de las líneas futuras de esta investigación.

REFERENCIAS BIBLIOGRÁFICAS

1. Stockley I. Interacciones Farmacológicas. Barcelona: Pharma Editores; 2004.

2. Amariles P, Giraldo NA, Faus MJ. Interacciones medicamentosas: aproximación para establecer y evaluar su relevancia clínica. Med Clín. 2007;129(1):27 -35.

3. Rodríguez-Terol A, Santos-Ramos B, Caraballo-Camacho M, Ollero-Baturone M. Relevancia clínica de las interacciones medicamentosas. Med Clín. 2008;130(19):758-59.

4. Thomson Healthcare Micromede x. 2012 [citado: 13-07-2012]. Disponible en: http://www.micromedex.com

5. Lexi-Comp, Inc. Lexi-interact. 2012 [consultado: 11-07-2012]. Disponible en: http://www.lexi.com

6. Minh VL, McCart GM, Tsourounis C. An assessment of free, online drug -drug interaction screening programs (DSPs). Hospi tal Pharmacy. 2003;38(7):662-68.

7. Hansten PD, Horn JR. Drug Interactions Analysis and Management. St. Louis: Facts and Comparations;2007.

8. Rodríguez-Terol A, Caraballo M, Palma D, Santos-Ramos B, Molina T, Desongles T, Aguilar A. Calidad estructural de las bases de datos de interacciones. Farm Hosp. 2009;33(3):134-46.

9. Cunningham H. Information Extraction, Automatic. Enclyclopedia of Language and Linguistics. Oxford: Elsevier;2005.

10. Proux D, Rechenmann F, Julliard L. Detecting Gene Symbols and Name s in Biological Texts: a First Step toward Pertinent Information Extraction. Proceedings of Genome Informatics. 1998;78-80.

11. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. Pr oceedings of the Pacific Symposium on Biocomputing. 2000;5:538-49.

12. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural -language processing system for the extraction of molecular pathways from journal articles. Bioinformatics. 2001;17(1):74-82.

13. Hirschman L, Park C, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. Bioninformatics. 2002;18(12):1553 -61.

14. Hearst M. Untangling text data mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistic. 1999;3 -10.

15. United States Adopted Names Council [citado: 12-07-2012]. Disponible en: http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council/naming-guidelines/approved-stems.page?

16. Jurafsky D, Martin J. Speech and language processing. A n introduction to natural language processing, Computational linguistics, and speech recognition. New Jersery: Prentice-Hall; 2000.

17. Karttunen L. Constructing lexical transducers. Proceedings of the 15th conference on Computational linguistics. Kyoto: C oling 94.1994;406-11.

18. Rodríguez S, Carretero J. A formal approach to Spanish morphology: the COES tools. XII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Sevilla: SEPLN. 1996;118-26.

19. Siddiqui T, Tiwary US. Natural language processing and information retrieval. New Dehli: Oxford University Press; 2008.

20. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics. 1995;21(4):543-65.

21. Lavid J. Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI. Madrid: Cátedra; 2005.

22. Rodríguez H. Técnicas básicas en el tratamiento informático de la lengua. Quark. Ciencia, Medicina, Comunicación y Cultura. 2000;19:26-34.

23. Johnson CD. Formal aspects of phonological description. La Haya: Mouton; 1972.

24. Koskenniemi K. Two-level morphology: a general computational model for word-form recognition and production. University of H elsinki: Department of General Linguistics; 1983.

25. Hopcroft JE, Ullman JD. Introduction to Automata Theory, Languages and Computation. Reading, MA: Addison-Wesley;1979.

26. Karttunen L, Kaplan RM, Zaenen A. Two-level morphology with composition. Proceedings of the 15th International Conference on Computational Linguistics. Nantes, France: Coling 92. 1992.

27. Abney S. Partial parsing via finite-state cascades. Journal of Natural Language Engineering. 1996;2(4):337-44.

28. Roche E, Schabes Y. Deterministic part-of-speech tagging with finite state transducers. Computational Linguistics. 1995;21(2):227-53.

29. Silberztein M. NooJ Manual. 2002 [citado: 14-07-2012]. Disponible en: http://www.nooj4nlp.net

30. _____. Complex Annotations with NooJ. Proceedings of the 2007 International NooJ Conference. Newcastle: Cambridge Scholars Publishing. 2008:214-27.

31. Wilbur WJ, Hazard GF, Divita G, Mork JG, Aronson AR, Browne AC. Analysis of biomedical text for chemical names: a comparison of three methods Proceedings AMIA Annual Symposium. 1999;176-80.

32. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: Extraction of drugs, genes and Relations from the biomedical Literature. Pacific Symposium on Biocomputing. 2000;5:514-25.

33. Segura Bedmar I, Martínez P, Samy D. Detección de fármacos genéricos en textos biomédicos. Procesamiento del Lenguaje Natural. 2008;40:27-34.

34. Hersh WR, Bhupatiraju RT. TREC genomics track overview, The Twelfth Text Retrieval Conference - TREC 2003;14-23.

35. Hersh W, Bhupatiraju RT, Ross L, Johnson P, Cohen AM, Kraemer DF. TREC 2004 genomics track overview. The Thirteenth Text Retrieval Conference - TREC 2004;13-24.

36. Tanabe L, Xie N, Thom LH, Matten W and Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics. 2005;6(Suppl. 1):S3.

37. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. BMC Bioinformatics. 2005;6:S1.

Recibido:14 de mayo de 2012.
Aceptado: 23 de julio de 2012.

Prof. *Carmen Gálvez***.** Departamento de Información y comunicación. Universidad de Granada. Campus Cartuja 18071, Granada, España. Correo electrónico: cgalvez@ugr.es