Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 65-72 recibido 30-03-2015 revisado 27-04-2015 aceptado 18-05-2015

Estudio de fiabilidad y viabilidad de la Web 2.0 y la Web semántica para enriquecer lexicones en el dominio farmacológico*

Web 2.0 and Semantic Web Reliability and Viability Study to Enhance Lexicons for the Pharmacological Domain

Isabel Moreno
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
imoreno@dlsi.ua.es

Paloma Moreda
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
moreda@dlsi.ua.es

M. Teresa Romá-Ferri
Dpt. Enf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
mtr.ferri@ua.es

Resumen: Los actuales sistemas de Reconocimiento de Entidades en el dominio farmacológico, necesarios como apoyo para el personal sanitario en el proceso de prescripción de un tratamiento farmacológico, sufren limitaciones relacionadas con la falta de cobertura de las bases de datos oficiales. Parece por tanto necesario analizar la fiabilidad de los recursos actuales existentes, tanto en la Web Semántica como en la Web 2.0, y determinar si es o no viable utilizar dichos recursos como fuentes de información complementarias que permitan generar y/o enriquecer lexicones empleados por sistemas de Reconocimiento de Entidades. Por ello, en este trabajo se analizan las principales fuentes de información relativas al dominio farmacológico disponibles en Internet. Este análisis permite concluir que existe información fiable y que dicha información permitiría enriquecer los lexicones existentes con sinónimos y otras variaciones léxicas o incluso con información histórica no recogida ni mantenida en las bases de datos oficiales.
Palabras clave: Reconocimiento de Entidades Nombradas; Farmacología; Lexicones; Enriquecimiento; Web 2.0; Web Semántica

Abstract: Nowadays Named Entity Recognition systems in the pharmacological domain, which are needed to help healthcare professional during pharmacological treatment prescription, suffer limitations related to the lack of coverage in official databases. Therefore, it seems necessary to analyse the reliability of existing resources, both in the Semantic Web and Web 2.0, and determine whether it is feasible or not to use these resources for additional information to generate and/or enhance lexicons used by Named Entity Recognition systems. For this reason, this paper analyses the main sources of information related to the pharmacological domain available on the Internet. This analysis leads to the conclusion that there is reliable information and it would enhance existing lexicons with synonyms, variations and even historical information not collected or maintained in official databases.
Keywords: Named Entity Recognition; Pharmacology; Lexicons; Enhancement; Web 2.0; Semantic Web

1 Introducción

Hoy en día disponemos de una gran cantidad de información digital relativa a la salud. Dicha información, en su mayoría textual, se encuentra disponible en fuentes de información heterogéneas como bases de datos o enciclopedias. Emplear toda esta información resulta crítico en el ámbito sanitario (Friedman, Rindflesch, y Corn, 2013). Por ejemplo, en varios estudios se pone de manifiesto que, para el personal sanitario, la prescripción de un tratamiento farmacológico es una situación crítica y frecuente (Ely et al., 1999; Gonzalez-Gonzalez et al., 2007). La prescripción está relacionada con la selección adecuada de los medicamentos (nombre identificativo con el que se comercializan) y de sus principios activos (los componen-

* Este trabajo ha sido financiado parcialmente por la Secretaria de Estado de Investigación, Desarrollo e Innovación - Ministerio de Economía y Competitividad (TIN2012-38536-C03-03 y TIN2012-31224)

ISSN 1135-5948

tes que aportan las cualidades al medicamento). El hecho de poder consultar diferentes fuentes de información ayudaría a los profesionales en este proceso de toma de decisiones. Sin embargo, acceder y analizar toda la información textual disponible resulta: (i) inmanejable para los profesionales sanitarios (Gonzalez-Gonzalez et al., 2006); y (ii) difícil de procesar por procesos automáticos (Meystre et al., 2010; Friedman, Rindflesch, y Corn, 2013). Además, es importante destacar que cualquier fuente de información digital no oficial, requiere un proceso de validación y verificación que determine su fiabilidad.

Una línea de investigación que aborda los obstáculos aquí expuestos es el Procesamiento del Lenguaje Natural (PLN). Su finalidad es proporcionar los mecanismos necesarios para convertir la información textual, fácil de comprender por humanos, en una representación comprensible para procesos computacionales, sin importar su volumen (Friedman, Rindflesch, y Corn, 2013). Para nuestros fines, entre las diferentes tareas que se engloban dentro del PLN destaca la tarea denominada Reconocimiento de Entidades Nombradas (REN). Dicha tarea tiene como objetivo identificar aquellos elementos de información relevantes en un texto y asignarles una categoría, de entre un conjunto predefinido, para su clasificación (Feldman y Sanger, 2007). En el ámbito sanitario, y en concreto durante la prescripción de un tratamiento farmacológico, ejemplos de estas categorías podrían ser los medicamentos y los principios activos.

Como se mostró en Moreno, Moreda, y Romá-Ferri (2015), para lograr su finalidad muchos sistemas REN se apoyan en lexicones especializados, los cuales se componen de un listado de términos que representan el vocabulario habitual para cada una de las categorías predefinidas del sistema. Para que estos sistemas resulten de ayuda en el dominio farmacológico, es muy importante que los términos incluidos en estos repositorios se obtengan de fuentes fidedignas. Un ejemplo de fuente fiable para lengua castellana es la base de datos Nomenclator de Prescripción[1], donde podemos encontrar todos los medicamentos autorizados, suspendidos y revocados en España a partir de mayo de 2013, así como los principios activos que los componen.

Una particularidad de este dominio es su evolución constante, causada por el descarte o la introducción de nuevos medicamentos autorizados en cada país. Por ello, aunque la fuente de información empleada para crear el lexicón sea fiable, los sistemas REN se encuentran con una serie de obstáculos relacionados, principalmente, con su cobertura. Uno de los problemas más importantes es la carencia de sinónimos y variantes léxicas (como plurales o abreviaturas), así como una cobertura temática reducida a los términos empleados en España, en el caso de Nomenclator. Estas limitaciones influyen en los resultados que puede alcanzar un sistema REN farmacológico diseñado para el procesamiento de información en castellano (Moreno, Moreda, y Romá-Ferri, 2015). Como consecuencia, es necesario buscar fuentes de información complementarias que nos permitan superar estos problemas, sin afectar a la fiabilidad de los lexicones. Actualmente Internet proporcion una gran variedad de recursos con información farmacológica de interés. El trabajo pendiente es analizar tales recursos y determinar o no su fiabilidad a la hora de ser utilizados en tareas de REN. Por ello, el objetivo de este artículo es estudiar fuentes de información alternativas y disponibles en la red, y averiguar si permiten ampliar los lexicones creados a partir de la información disponible en Nomenclator, sin perder por ello fiabilidad.

El resto del artículo está organizado como sigue. La sección 2 describe y caracteriza las fuentes de información analizadas en este trabajo. A continuación, la sección 3 detalla cómo se ha obtenido la información de la fuente seleccionada. Después, en la sección 4, proponemos un método de validación automático para ayudar a un experto en el análisis manual. Seguidamente, en la sección 5, se describe dicho análisis manual. Terminamos con las conclusiones y el trabajo futuro en la sección 6.

2 Recursos de la Web 2.0 y la Web Semántica

En la red podemos encontrar diversos recursos o bases de conocimiento con información farmacológica de interés. Estas fuentes de información siguen dos filosofías diferentes: (i) la Web 2.0, destinada a los humanos, está organizada de forma semiestructurada, es de-

[1] `http://www.aemps.gob.es/cima/pestanias.do?metodo=nomenclator` (Último acceso: 13 Febrero 2015)

Nombre	Estructura	Tipo	Creación	Fuentes	NPA
Wikcionario	SE	D	MC	-	9
Wikipedia	SE	EN	MC	-	35
BabelNet	E	B	A	WordNet, Wikipedia, etc.	717
Wikidata	E	B	MC	-	492
DBpedia	E	B	SA	Wikipedia	921

Acrónimos: (i) A: Automático; (ii) B: Base de datos; (iii) D: Diccionario; (iv) E: Estructurado; (v) EN: Enciclopedia; (vi) MC:Manual y Colaborativo; (vii) NPA: Número de instancias clasificadas como Principo Activo; (viii) SA: Semi-Automático; (ix) SE: Semi-Estructurado.

Tabla 1: Fuentes de información de la Web 2.0 y la Web Semántica en castellano

cir, contiene información textual sin una estructura bien definida; mientras que (ii) la Web Semántica, destinada tanto a usuarios como a procesos automáticos, organiza su conocimiento de forma estructurada, es decir, la información textual contiene metadatos que facilitan los procesos automáticos. Concretamente, hemos analizado si las cinco fuentes de información más utilizadas en el dominio general, nos permitirían obtener términos que identifiquen medicamentos y principios activos, para su posterior inclusión en un lexicón para lengua castellana. En la tabla 1 encontramos un resumen de este análisis.

- Wikcionario[2]: diccionario multilinngüe colaborativo. Es un recurso semiestructurado, donde cada palabra tiene varias secciones. Para este estudio, destaca la sección de variantes pues incluye otras formas léxicas para cada entrada. Además, cada entrada puede estar asignada a una o varias categorías. A su vez, cada categoría puede dividirse en otras categorías más específicas. Estas categorías permiten seleccionar aquellas entradas del diccionario clasificadas como un principio activo: "*Categoría:ES:Fármacos*"[3], formada por 9 términos en castellano. No incluye entradas sobre medicamentos.
- Wikipedia[4]: enciclopedia multilingüe colaborativa con una gran cobertura. Cada artículo en Wikipedia está parcialmente estructurado, es decir, en su mayoría es texto libre que consta de información estructurada como las categorías y las cajas de información (ver Figura 1). En concreto, los principios activos se encuentran en la "*Categoría:Fármacos*", la cual se compone de 35 artículos y 15 subcategorías, que a su vez pueden contener más artículos y más subcategorías, en castellano. De cada caja de información, se puede obtener datos como: su nombre químico normalizado establecido por IUPAC[5] (International Union of Pure and Applied Chemistry), o su clasificación ATC[6] (Anatomical Therapeutic Chemical classification, sistema europeo de codificación de sustancias farmacéuticas y medicamentos). Dichos datos permitirían contrastar la fiabilidad de la información disponible. De nuevo, no existe una categoría que represente a los medicamentos.
- BabelNet[7](Navigli y Ponzetto, 2012): red semántica multilingüe cuyo propósito es ofrecer un diccionario enciclopédico combinando WordNet(Miller et al., 1990) y Wikipedia. Cada entrada en esta red contiene información estructurada, incluyendo un conjunto de definiciones en varios idiomas y tanto su categoría gramatical como su categoría en Wikipedia. No incluye medicamentos pero sí principios activos en español. Sin embargo, no se clasifican con la misma categoría descrita para Wikipedia, sino que usan va-

[2] `https://es.wiktionary.org` (Último acceso: 9 Marzo 2015)

[3] `https://es.wiktionary.org/wiki/Categor\%C3\%ADa:ES:F\%C3\%A1rmacos` (Último acceso: 21 Marzo 2015)

[4] `https://es.wikipedia.org` (Último acceso: 9 Marzo 2015)

[5] `http://www.iupac.org/` (Último acceso: 21 Marzo 2015)

[6] `http://www.whocc.no/atc/structure_and_principles/` (Último acceso: 21 Marzo 2015)

[7] `http://babelnet.org/` (Último acceso: 21 Marzo 2015)

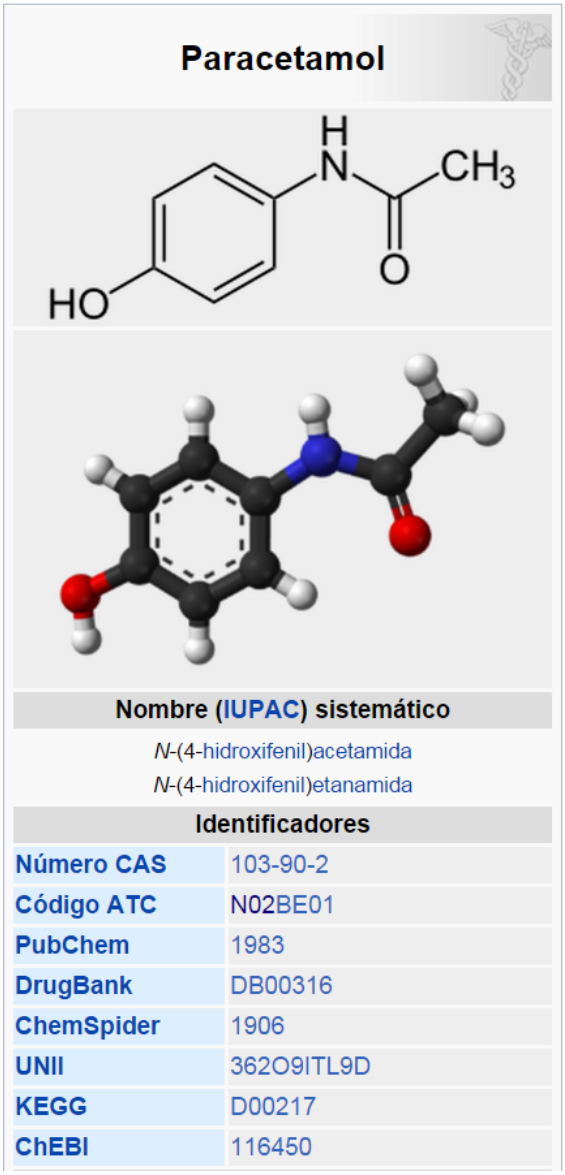

Figura 1: Fragmento de una caja de información de Wikipedia

rias subcategorias relacionadas con el nivel ATC al que el principio activo pertenece (por ejemplo:'*Categoría:Código_ATC_A*"). Contiene 717 principos activos en castellano.

- Wikidata[8](Vrandečić y Krötzsch, 2014): base de datos colaborativa, cuyas propiedades pueden ser leídas y editadas tanto por humanos como por máquinas. Proporciona una fuente común para datos factuales en Wikipedia,esto es cada caja de información en una página de Wikipedia es una entrada en Wikidata. En particular, los principios activos se encuentran en la categoría de Wikipedia "*Compuesto químico*" y se incluyen 492 instancias, independientemente del idioma. Contiene información complementaria como la vía de administración y su fórmula química. Los nombres de los medicamentos disponibles se tratan como sinónimos del principio activo. Por ello no hay manera de distinguir si Wikidata nos puede ofrecer principios activos o medicamentos.

- DBpedia[9](Lehmann et al., 2012): base de conocimiento multilingüe que extrae la información estructurada de Wikipedia. Cada recurso se mapea a una página de Wikipedia basándose en su título. Esta base de conocimiento se construye usando varios procedimientos de extracción. Uno de ellos es definir manualmente mapeos que relacionen las cajas de información de Wikipedia con la ontología DBpedia, produciendo así datos de mayor calidad. La versión española no incluye una clasificación para principios activos ni para medicamentos. Por el contrario, la versión inglesa dispone de una categoría para principios activos, "*dbpedia-owl:Drug*", aunque no de una categoría para medicamentos. Además, esta versión incluye los nombres de principios activos tanto en castellano como en inglés. De esta ontología, al igual que de Wikipedia, se pueden consultar una serie de propiedades, como el nombre de la IUPAC y su clasificación ATC, lo que permitiría contrastar su fiabilidad para los principios activos candidatos que incluye. Contiene 921 recursos clasificados como principio activo tanto en inglés como en castellano.

Tras esta revisión observamos que: (i) la mayoría de estos recursos se basan en la Wikipedia, por lo que la información facilitada puede ser fácilmente contrastada a través de los códigos ATC; (ii) todos ellos incluyen una categoría para el concepto de principio activo pero ninguno contempla el concepto medicamento; (iii) la fuente de información que contiene más principios activos actualmente es DBpedia; (iv) el acceso a DBpedia es sencillo

[8]`http://www.wikidata.org/?uselang=es` (Último acceso: 9 Marzo 2015)

[9]`http://dbpedia.org/` (Último acceso: 9 Marzo 2015)

puesto que la información está estructurada; y (v) aunque para poder acceder a la información sobre principios activos es necesario hacerlo a través de la versión en inglés de DBpedia, ésta aporta la información también en castellano. Por ello, en este trabajo nos centraremos en analizar si los principios activos incluidos en DBpedia son fiables y si es viable o no emplearlos como fuente única o bien como fuente complementaria en la creación de lexicones.

3 Obtención de principios activos de DBpedia

Como se ha comentado previamente, para obtener los principios activos de DBpedia es necesario acceder a la versión inglesa pues dispone de la categoría semántica principio activo ("*dbpedia-owl:Drug*") y, además, éstos pueden obtenerse en castellano mediante una etiqueta identificativa del idioma (séptima línea de la Figura 2). Dado que la información está estructurada, puede obtenerse fácilmente una lista de principios activos realizando una consulta sobre la base de conocimiento. Como se observa en la Figura 2, se ha empleado el lenguaje de consulta SPARQL[10] para recuperar: (i) la URI del recurso de DBpedia que representa un principio activo (por ejemplo, el principio activo "paracetamol" tiene la URI: "`http://dbpedia.org/resource/Paracetamol`"); (ii) su nombre ("*rdfs:label*") en español (ver las líneas sexta y séptima de la Figura 2); junto con (iii) su código ATC, el cual está divido en dos partes, la parte inicial del código ("*dbpedia-owl:atcPrefix*") y la parte final ("*dbpedia-owl:atcSuffix*"). Se han incluido aquellos principios activos cuyo código de la ATC ("*dbpedia-owl:atcPrefix*") comenzaba entre las letras A y V (ver la octava línea de la Figura 2) por dos razones: (i) son sustancias que en algún momento se han empleado para la prevención, curación o mejora de una enfermedad sufrida por un humano; y (ii) forman parte de la ATC de la Organización Mundial de la Salud (OMS). Ambas, nos aportan la confianza de un vocabulario estándar, lo que nos permitirá comprobar la validez de los candidatos obtenidos.

Como resultado de esta consulta SPARQL, hemos recuperado 921 nombres de principios activos en español con sus correspondientes códigos ATC. El siguiente paso será determinar la fiabilidad de la información obtenida. Este proceso de validación se detalla en la Sección 4.

[10] `http://dbpedia.org/sparql` (Último acceso: 14 Marzo 2015)

```
select ?ppio, (CONCAT(?prefijo, ?sufijo) AS ?codigo), ?nombre where
 {
   ?ppio rdf:type dbpedia-owl:Drug .
   ?ppio dbpedia-owl:atcPrefix ?prefijo.
   ?ppio dbpedia-owl:atcSuffix ?sufijo.
   ?ppio rdfs:label ?nombre.

   FILTER(langMatches(lang(?nombre), "ES"))
   FILTER(regex(?prefijo,"^[A-V]"))
 }
order by ?prefijo ?sufijo
```

Figura 2: Consulta SPARQL para recuperar principios activos de DBpedia

4 Validación automática de los principios activos recuperados de DBpedia

Cada una de los 921 instancias relativas a principios activos obtenidas de DBpedia está compuesta por su código ATC y su nombre. Para contrastar la validez de estos principios activos hemos establecido un procedimiento de comparación automático en dos fases (Secciones 4.1 y 4.2). El objetivo de dicha comparación es seleccionar aquellos principios activos o variaciones sobre ellos no existentes en el lexicón ActILex (Moreno, Moreda, y Romá-Ferri, 2015), generado a partir de Nomenclator, y por tanto, candidatos a ser incluidos como complemento. En la primera fase, denominada filtrado de códigos, se comparan los códigos ATC obtenidos de DBpedia con los de fuentes fiables con el fin de identificar principios activos no incluidos en las bases de datos oficiales. En la segunda, denominada comparación de términos, se intentan comparar nombres de principios activos para aquellos casos en los que no existe coincidencia en el código. Como resultado de este proceso se obtendrá la lista de principios activos candidatos a enriquecer ActILex y que habrán por tanto de ser validados por un experto del dominio.

4.1 Filtrado de códigos

En este paso se ha realizado una comparación entre los códigos ATC recuperados de DBpedia y los códigos ATC del lexicón ActILex (Moreno, Moreda, y Romá-Ferri, 2015). Este lexicón se genera a partir de Nomenclator, una base de conocimiento oficial. Incluye los principios activos (nombre y código) de la versión 03 2011 eliminando aquellas entradas cuyos códigos ATC comenzaban por W, X,

Y y Z, dado que especifican productos sanitarios de uso exclusivo en el sistema de atención español. A la versión actual, además, se le añadieron nuevos códigos y nombres de principios activos de la versión Nomenclator 20-02-2015, siguiendo así la misma orientación establecida en la sección 3 para DBpedia.

Cuando los códigos ATC son iguales en ambos recursos, nos encontramos ante un principio activo válido que no requiere de ningún proceso de verificación por parte de expertos. En concreto, obtenemos 789 códigos ATC coincidentes, de los 921.

4.2 Comparación de términos

Los 132 principios activos para los que no se encontró coincidencia en la fase anterior a través del código ATC, son considerados ahora por su nombre. Para ello, primero se emplea el buscador del Centro de Información online de Medicamentos[11](CIMA) de la AEMPS para verificar automáticamente tanto el nombre como el código del principio activo. Cuando los códigos eran coincidentes, se ha verificado si el nombre en ambos recursos era exactamente el mismo. En caso de encontrar coincidencia el principio activo era considerado válido y eliminado de la lista de elementos a analizar manualmente por un experto tal y como se detalla en la Sección 5.

Asimismo, un proceso automático ha confirmado si el mismo nombre de principio activo de DBpedia se encontraba en ActILex, pero con otro código. Cuando los nombres eran coincidentes, el principio activo era considerado válido y se eliminaba de la lista de elementos a verificar por un experto.

Como resultado de este proceso sólo 69 del conjunto inicial de principios activos quedaron en la lista de elementos a validar por un experto.

5 Análisis de fiabilidad de los principos activos extraídos de DBpedia

El proceso de comparación automático presentado en la sección anterior dio lugar a un conjunto de principios activos cuyo código y término de identificación no coincidía con ninguna de las fuentes de referencia consultadas, por lo que se determinó la comprobación manual por un experto. Dicha revisión se centro en determinar:

(i) Grupo de principios activos que no coincide ni con el código ni con el término de identificación. Tras la comprobación manual se encontró que 9 principios activos localizados en DBpedia no deberían ser considerados como tales. Uno de ellos por tratarse de una sustancia no reconocida por la OMS para uso generalizado (el caso de Picamilón, http://dbpedia.org/resource/Picamilon).
Los restantes 8 principios activos se descartaron por ser sustancias pertenecientes a la clasificación ATC veterinaria[12]. Cabe mencionar que estas sustancias en la versión inglesa de Wikipedia, fuente de origen para DBpedia, se identifican como tales principios activos veterinarios con su código característico, iniciado por la letra "Q". Sin embargo, en la carga de información estructurada de DBpedia esta parte inicial del código no es incluida. Un ejemplo de esta situación es el principio activo Sulfadoxina (http://dbpedia.org/resource/Sulfadoxine), en DBpedia se le asigna el código J01EQ13 mientras que su código real es el QJ01EQ13, disponible en Wikipedia.
No obstante, el experto confirmó que 4 de los principios activos sí que eran válidos, para ello utilizó otras fuentes de información, tanto de caracter nacional (Vademecum) como internacional (PubChem Classification Browser). Estos principios activos son: nafcilina, fenibut, carfilzomib y lorcaserina.

(ii) En la comprobación de principios activos con códigos iguales se detectaron aquellos con un término de identificación diferente al proporcionado por CIMA. A este nivel se localizaron 25 principios activos de DBpedia que eran o bien sinónimos o variaciones léxicas de los nombres contenidos en el lexicón propio. En concreto, el 64 % eran sinónimos (ejemplo: "Ácido glicirrícico", en CIMA y "glicirricina", en DBpedia) y el 36 % eran variantes léxicas (ejemplo de género: "cinoxacinO", en CIMA, y "cinoxacinA", en DBpedia).

(iii) El último conjunto de resultados corresponde a 24 principios activos cuyo código ATC no coincide con ActILex. En este caso se comprobó la coincidencia del término de identificación del principio activo con los términos del lexicón ActILex. La premisa que

[11] `http://www.aemps.gob.es/cima/pestanias.do?metodo=accesoAplicacion` (Último acceso: 14 Marzo 2015)

[12] `http://www.whocc.no/atcvet/atcvet/` (Último acceso: 14 Marzo 2015)

sustenta este tipo de comprobación se basa en que un término de un principio activo es asociado a uno o varios códigos, de acuerdo a su funcionalidad terapéutica. No obstante, cuando se demuestra que una funcionalidad terapéutica no es adecuada el código es eliminado; en este caso, el término sigue en vigor pero perdiendo uno de sus código de identificación. La comprobación confirmó que los 24 términos coincidían de forma completa con términos acumulados en nuestro lexicón de principios activos.

Atendiendo al análisis realizado, se puede concluir que la información facilitada por DBpedia tiene utilidad para enriquecer un lexicón especializado en principios activos en castellano empleados por sistemas REN, puesto que incluiría sinónimos, variantes léxicas y códigos obsoletos. Sin embargo, DBpedia no debería ser empleado como la única fuente para un sistema REN basado en diccionarios debido a las limitaciones que presenta, ya que por ejemplo contiene principos activos de uso veterinario.

6 Conclusiones y trabajos futuros

Este trabajo ha realizado un análisis relativo a la viabilidad de utilizar o no fuentes de información disponibles en la red y procedentes de fuentes no oficiales, para la generación y/o el enriquecimiento de lexicones que ayuden en la tarea de REN en el dominio farmacológico. Dicho análisis ha permitido establecer que si bien la generación no es fiable, puesto que contiene sustancias de uso no generalizado por la OMS y principios activos de uso veterinario; el enriquecimiento sí que sería deseable, ya que aporta sinónimos y variaciones léxicas, así como de términos de identificación de principios activos obsoletos. Estos últimos no deben de ser ignorados (Cimino, 1998), sino que deben incluirse como datos históricos para permitir así su detección, pues en algún momento fueron usadas para los tratamientos farmacoterapéuticos. El estudio ha quedado reducido a los principios activos puesto que ninguno de los recursos considerados ofrecía información sobre medicamentos.

A pesar de ello, la mejora relativa a principios activos, justifica el estudio y plantea como trabajo futuro la necesidad de definir el proceso que permita incorporar de forma automática los elementos no contenidos en el lexicón ActILex, generado a partir de Nomenclator, y enriquecerlo con sinónimos, variantes léxicas y entidades obsoletas procedentes de DBpedia. Así como cuantificar con métodos estadísticos la aportación de estas nuevas entradas en la efectividad de un sistema REN basado en diccionarios, lo que permitirá confirmar si el aumento de cobertura es significativo al emplear la versión enriquecida de ActILex.

Bibliografía

Cimino, J. J. 1998. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine - Author manuscript; available in PubMed Central 2012 August 10*, 37(4-5):394–403.

Ely, J. W, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T Levy, M. Chambliss, y E. R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, 319:358–361.

Feldman, R. y J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, New York, 2009 edición.

Friedman, C., T. C. Rindflesch, y M. Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73, Octubre.

Gonzalez-Gonzalez, A. I., M. Dawes, J. Sanchez-Mateos, R. Riesgo-Fuertes, E. Escortell-Mayor, T. Sanz-Cuesta, y T. Hernandez-Fernandez. 2007. Information Needs and Information-Seeking Behavior of Primary Care Physicians. *Annals of Family Medicine*, 5:345–352.

Gonzalez-Gonzalez, A. I., J.F Sanchez Mateos, T. Sanz Cuesta, R. Riesgo Fuertes, E. Escortell Mayor, y T. Hernandez Fernandez. 2006. Estudio de las necesidadesde información generadas por los médicos de atención primaria (proyecto ENIGMA)*. *Atención primaria*, 38(4):219–224.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mende, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Ex-

tracted from Wikipedia. *Semantic Web*, 1:1–5.

Meystre, S. M., J. Thibault, S. Shen, J. F. Hurdle, y B. R. South. 2010. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association*, 17(5):559–562.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K.J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Moreno, I., P. Moreda, y M. T. Romá-Ferri. 2015. MaNER: a MedicAl Named Entity Recogniser for Spanish. En *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015*.

Navigli, R. y S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Vrandečić, D. y M. Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85.