View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional de la Universidad de Alicante

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 83-90 recibido 25-03-2015 revisado 24-04-2015 aceptado 19-05-2015

Extracción no supervisada de relaciones entre medicamentos y efectos adversos*

Unsupervised extraction of adverse drug reaction relationships

Andrés Duque
NLP Group at UNED
28040 Madrid, Spain
aduque@lsi.uned.es

Juan Martínez-Romo
NLP Group at UNED
28040 Madrid, Spain
juaner@lsi.uned.es

Lourdes Araujo
NLP Group at UNED
28040 Madrid, Spain
lurdes@lsi.uned.es

Resumen: En este trabajo se presentan los resultados preliminares de una nueva técnica no supervisada para la extracción de relaciones entre medicamentos y efectos adversos. La identificación de relaciones se consigue a partir de un modelo de representación de conocimiento que extrae pares de entidades con un peso determinado, en función de la significatividad estadística de su coaparición en un mismo documento. Dicho modelo puede ser posteriormente convertido en un grafo. El sistema ha sido evaluado sobre un corpus de referencia, denominado ADE corpus, consiguiendo resultados prometedores al obtener una eficacia muy por encima de un *baseline* estándar. Las primeras pruebas también muestran un alto potencial para inducir conocimiento nuevo.
Palabras clave: Extracción de información, Dominio médico, Extracción de relaciones, Reacciones adversas a medicamentos

Abstract: In this work we present preliminary results obtained by a new unsupervised technique for extracting relations between drugs and adverse drug reactions. The identification of those relations is achieved using a knowledge representation model that generates pairs of entities and assigns them a specific weight, depending on the statistical significance of their co-occurrence in the same document. This model may subsequently be transformed into a graph. The system has been evaluated over the reference ADE corpus, obtaining promising results, since its effectiveness is quite higher than that obtained by a standard basline. First tests also show a high potential for inducing new knowledge.
Keywords: Information extraction, Medical domain, Relation extraction, Adverse drug effect

1. Introducción

La identificación de reacciones adversas a medicamentos (*ADR: Adverse Drug Reaction*) es una problemática de gran relevancia en la práctica médica. ADR (Edwards y Aronson, 2000) se define como cualquier forma nociva, no intencionada, de reacción no deseada o desagradable que resulta del uso de una dosis de un medicamento para el propósito de la profilaxis, el diagnóstico o la terapia. Predice el peligro de la futura administración y establece la necesidad del cambio de la dosis o de la retirada del producto. Habitualmente los efectos adversos de los medicamentos no se conocen por completo en el momento de su aprobación ya que los ensayos clínicos previos son de tamaño limitado y se realizan en un período de tiempo corto. Por ello es frecuente que posteriormente aparezcan efectos adversos adicionales, en algunos casos graves. Esto hace que sea de vital importancia monitorizarlos y reportarlos en el menor tiempo posible. La extracción automática de información es por tanto de gran ayuda en este proceso, ya que puede aliviar notablemente el trabajo manual, y se está explorando, tanto en documentos científicos (Gurulingappa et al., 2012) e informes clínicos (Aramaki et al., 2010), como en información extraída de sitios web (Segura-Bedmar, de la Peña González, y Martínez, 2014).

La extracción de relaciones en general y en este caso en particular requiere realizar dos tareas. En primer lugar es necesario identificar en el texto las entidades entre las que se pueden dar las relaciones buscadas. Posteriormente se trata de identificar los casos en los que se cumple la relación entre dos entidades. En los últimos años se han aplicado técnicas de procesamiento del lenguaje natural a ambos aspectos del problema, aunque nosotros nos centramos en el segundo, en el contexto de la documentación científica.

La identificación de relaciones se ha abordado con diversas técnicas. Algunas propuestas se ba-

* Trabajo financiado parcialmente por los proyectos EXTRECM (TIN2013-46616-C2-2-R), y TwiSE (2013-025-UNED-PROY).

ISSN 1135-5948
© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

san en la coaparición de las entidades de interés (Pyysalo et al., 2008; Kandula y Zeng-Treitler, 2010). En estos trabajos se supone que dos entidades que se mencionan en la misma frase o en el mismo resumen pueden estar relacionadas. Lógicamente, este enfoque proporciona una cobertura alta, pero una precisión muy baja. Debido a su sencillez se suele adoptar como *baseline* para hacer comparativas con otros métodos. En Wang et al. (2009) el sistema MedLEE es aplicado para identificar potenciales ADRs en resúmenes. Realizan pruebas basadas en la distribución χ^2 para seleccionar las asociaciones. Por su parte, el sistema descrito en Kang et al. (2014) utiliza una base de conocimiento para la identificación de relaciones. Concretamente se usa una representación en forma de grafo de la información contenida en el metatesauro UMLS (Lindberg, Humphreys, y McCray, 1993). UMLS define términos y conceptos, así como relaciones entre los conceptos. Estos autores utilizan distancias entre conceptos para seleccionar relaciones. Otros autores han utilizado un enfoque de aprendizaje automático (Gurulingappa et al., 2011; Gurulingappa, Mateen-Rajput, y Toldo, 2012). En (Eltyeb y Salim, 2015) se aplica un sistema basado en patrones para identificar las asociaciones. Los patrones se identifican automáticamente y después se pueden utilizar para aumentar la base de datos de relaciones. Sin embargo, no se han encontrado trabajos en la literatura que utilicen aproximaciones no supervisadas al problema.

En este trabajo nos centramos en la segunda fase de extracción de relaciones, en la que las entidades a relacionar ya han sido anotadas. En concreto, se aplica un refinamiento del modelo basado en co-ocurrencia. Como otros trabajos, suponemos que la coaparición de dos entidades en el mismo documento (en este caso, el resumen de un artículo médico) puede considerarse una indicación de una posible relación entre ellas. Sin embargo, sólo consideramos que la coaparición de dos entidades es representativa si su frecuencia es estadísticamente significativa respecto a la aparición de las entidades por separado.

El resto del artículo se organiza de la siguiente forma: en la Sección 2 se describe el corpus y la técnica utilizada para etiquetar los resúmenes. En la Sección 3 se detalla el proceso completo de extracción de relaciones, a través del modelo de representación del conocimiento propuesto. Las Secciones 4 y 5 se centran en la experimentación y el análisis de resultados. Finalmente, en la Sección 6 se extraen las principales conclusiones y se exponen las líneas de trabajo futuro.

2. Materiales y Métodos

2.1. Preparación del Corpus

Como base de conocimiento para nuestro sistema, se ha seleccionado el corpus ADE (Gurulingappa et al., 2012), que detalla relaciones entre medicamentos y efectos adversos extraídas a partir de un conjunto de 2972 resúmenes de artículos, almacenados en Medline. Del corpus inicial, construido con una base de 5063 medicamentos y 5776 condiciones médicas, son públicos un total de 1644 resúmenes, aquéllos que presentan al menos una frase describiendo un efecto adverso. En total el corpus contiene un total de 6821 relaciones de efecto adverso a un medicamento. El archivo que almacena dichas relaciones se va a utilizar en el presente trabajo como *Gold Standard* de relaciones entre medicamentos y efectos adversos, tras eliminar las relaciones repetidas. En dicho archivo, cada relación se expresa con una serie de campos, separados por el carácter "|" , que contienen respectivamente el identificador del resumen en la base de datos de Medline, la frase de la que se extrae la relación, el efecto adverso, las posiciones de inicio y final del efecto adverso en la frase, el medicamento, y las posiciones de inicio y final del medicamento en la frase. Por ejemplo, la línea *"3159106 |Allopurinol hypersensitivity. |hypersensitivity |21 |37 |Allopurinol |9 |20"* define el efecto adverso *"hypersensitivity"* provocado por el medicamento *"allopurinol"*, que puede ser encontrado en el resumen con identificador "3159106".

La Tabla 1 contiene los datos del corpus ADE original, así como los datos útiles del mismo, utilizados para construir el *Gold Standard* que se utiliza en el presente trabajo.

	Corpus	**Gold Standard**
Resúmenes	2972	1644
Medicamentos	5063	1049
Efectos Adversos	5776	2983
Relaciones	6821	5098

Tabla 1: Estadísticas del corpus ADE, en su versión original (columna **Original**), y tras extraer los elementos útiles de los resúmenes que contienen al menos una relación entre medicamento y efecto adverso.

Además del número de resúmenes, en la tabla se puede observar el número de medicamentos, efectos adversos, y relaciones no repetidas que se encuentran en el *Gold Standard*.

2.2. Anotación de los resúmenes

El corpus ADE ofrece los identificadores de Medline de los resúmenes que contienen relaciones entre medicamentos y efectos adversos. Sin embargo, para la aplicación de nuestro algoritmo es necesario etiquetar dichos resúmenes con todos los posibles medicamentos y efectos adversos que aparezcan en los mismos. Por tanto, se accede a los resúmenes vía PubMed, y se identifican sobre el texto de cada uno de dichos resúmenes aquellas entidades (medicamentos o efectos adversos) susceptibles de aparecer en una relación. Para ello, se utilizan las listas de medicamentos y efectos adversos extraídas del corpus ADE. Una vez hecho esto, cada documento quedará representado por una bolsa de entidades etiquetadas, que serán las que nos permitan elaborar el grafo de coaparición. El número total de entidades en el *Gold Standard* es de 13642, mientras que el número total de entidades etiquetadas es de 25687. Esto nos da una idea de la dificultad de encontrar las relaciones correctas de entre todas las posibles combinaciones entre entidades dentro de los documentos etiquetados.

Tal y como se ha adelantado en la Sección 1, es importante destacar que el objetivo fundamental de este trabajo es analizar la utilidad de la técnica basada en coaparición de entidades para la extracción de relaciones entre medicamentos y efectos adversos. Es decir, no nos centramos en la eficacia de una técnica de etiquetado concreta, sino que consideramos un etiquetado hipotéticamente perfecto, en el que se anotan todas las entidades que nos interesan, para analizar el comportamiento del sistema propuesto. Una técnica de etiquetado diferente introduciría un sesgo que es el que se pretende evitar a través del etiquetado propuesto.

3. Modelo de extracción de relaciones (significatividad estadística)

El siguiente paso consiste en el análisis de coaparición de entidades, a partir de los documentos etiquetados.

Para comprobar si la coaparición de dos entidades en un documento es significativa, se define un modelo nulo en el que las entidades se distribuyen aleatoria e independientemente entre un conjunto de documentos de un corpus. Concretamente, se calcula la probabilidad de que dos entidades coincidan por puro azar. Este valor nos permite determinar un p-valor p para la coaparición de dos entidades. Si $p \ll 1$ se puede considerar que la aparición de las dos entidades en el mismo documento es significativa, y por lo tanto, es probable que su significado esté relacionado.

Concretamente, si dos entidades se encuentran respectivamente en n_1 y n_2 documentos, de entre los N que componen el corpus, para contar cuantos casos existen en los que dos entidades coincidan en exactamente k documentos, debemos tener en cuenta que hay cuatro tipos de documentos: k documentos que contienen ambas entidades, $n_1 - k$ documentos que contienen sólo la primera entidad, $n_2 - k$ documentos que contienen sólo la segunda entidad, y $N - n_1 - n_2 + k$ documentos (siempre que este número no sea cero) que no contienen ninguna de las dos entidades. Por lo tanto, el número de disposiciones que buscamos viene dado por el coeficiente multinomial:

$$\binom{N}{k, n_1 - k, n_2 - k} \tag{1}$$

Así, la probabilidad de que dos entidades que aparecen en los documentos n_1 y n_2 respectivamente y que están distribuidas de forma aleatoria e independiente entre N documentos, coincidan en exactamente k de ellos viene dada por:

$$p(k) = \binom{N}{n_1}^{-1} \binom{N}{n_2}^{-1} \binom{N}{k, n_1 - k, n_2 - k}, \tag{2}$$

si $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ y cero en otro caso.

Podemos escribir la ecuación (2) de una forma más fácil de tratar computacionalmente. Para ello introducimos la notación $(a)_b \equiv a(a-1)\cdots(a-b+1)$, para cualquier $a \geq b$, y sin pérdida de generalidad suponemos que la primera entidad es la más frecuente, es decir $n_1 \geq n_2 \geq k$. Entonces:

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2} (k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2 - k} (N - n_2 + k)_k (k)_k}, \end{aligned} \tag{3}$$

donde en la segunda forma se ha usado la identidad $(a)_b = (a)_c (a-c)_{b-c}$ valida para $a \geq b \geq c$. La ecuación (3) se puede reescribir como

$$\begin{aligned} p(k) = &\prod_{j=0}^{n_2 - k - 1} \left(1 - \frac{n_1}{N - j}\right) \\ &\times \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)} \end{aligned} \tag{4}$$

Esto nos permite determinar un p-valor para la coaparición de dos entidades como

$$p = \sum_{k \geq r} p(k), \tag{5}$$

donde r es el número de documentos en el corpus en el que coaparecen las dos entidades.

La forma de proceder a partir de este punto es la usual en la comprobación de hipótesis estadísticas: se establece un nivel de confianza p_0 (habitualmente $p_0 \leq 0{,}05$, es decir, la hipótesis nula es incorrecta con un nivel de confianza del 95 % o superior) de manera que la coaparición es significativa sólo si $p < p_0$. De acuerdo con esto se define un par de entidades i y j relacionadas sólo si coaparecen de acuerdo con este criterio. Pero cuanto más bajo sea el valor de p más significativa es la coaparición, por lo que tiene sentido asignar un peso a esta relación. Esta significatividad se puede cuantificar tomando la mediana (correspondiente a $p = 1/2$) como una referencia y calculando el peso como $\ell = -\log(2p)$, es decir una medida de cuanto se desvía de la mediana el valor real de r (número de documentos en el corpus en los que coaparecen las dos entidades).

Mediante esta técnica se extrae un modelo de representación del conocimiento en el que se almacenan los pares de entidades que coaparecen de forma significativa. Estos pares de entidades se conectan con un peso que mide la significatividad de su coaparición. En este caso estamos interesados únicamente en aquellos pares de entidades que conecten medicamentos (entidades almacenadas con la etiqueta "MED") con efectos adversos (entidades almacenadas con la etiqueta "DIS"). De las entidades que están formadas por varias palabras se eliminan las denominadas *stopwords*, o palabras del propio idioma (en este caso inglés) que no aportan información.

4. Experimentación y resultados

La evaluación se ha realizado en términos de Precisión, Cobertura y Medida-F (*Precision*, *Recall* y *F-Measure*).

4.1. Baseline

Dentro de la evaluación de nuestro sistema, se ha desarrollado un *baseline*, obtenido mediante una técnica simple, que define un umbral de resultados a superar por el algoritmo propuesto. Dicho *baseline* se obtiene considerando que todas las entidades que aparecen en un mismo resumen están relacionadas, siempre que una de las entidades sea un medicamento y la otra un efecto adverso. El *Gold Standard* se construye a partir de frases que se encuentran en los resúmenes, es decir, toda relación que se encuentra en el *Gold Standard* se extrae de uno o varios documentos concretos. El *baseline* ofrece por tanto una cobertura perfecta, aunque su precisión es muy baja ya que contiene muchas relaciones incorrectas.

4.2. Resultados iniciales

La Tabla 2 muestra el *baseline*, así como los resultados obtenidos por la primera aproximación de nuestro algoritmo, descrito en la Sección 3.

Sistema	P	C	F
Baseline	25,20	**100,00**	40,25
Propuesto	**42,33**	59,67	**49,53**

Tabla 2: Resultados en función de la Medida-F (F), Precisión (P) y Cobertura (C) obtenidos por nuestro algoritmo, en comparación con el *baseline*. Los campos en negrita indican el mayor valor de cada medida.

Tal y como se indicaba anteriormente, el *baseline* consigue una cobertura perfecta del problema, sin embargo, la precisión (número de aciertos partido por el número total de relaciones propuestas) de nuestro algoritmo es mayor, lo que redunda en una mayor Medida-F. Es decir, aunque nuestro sistema no encuentra todas las relaciones posibles (encuentra alrededor de un 60 % de ellas), la proporción de aciertos es mayor, lo cuál nos indica que nuestro algoritmo está encontrando con mayor facilidad aquellas relaciones más significativas, desechando con efectividad otras que una técnica como la que implementa el *baseline* asignaría por el simple hecho de aparecer en un mismo documento. El umbral de significatividad estadística utilizado en nuestro algoritmo es de $P_0 = 0{,}01$, es decir, se aplica un nivel de confianza del 99 %.

4.3. Análisis de distancias

Tras los primeros resultados se realizó un análisis detallado de las relaciones obtenidas por el sistema. La primera impresión fue que existían numerosas relaciones redundantes, en comparación con aquéllas contenidas en el *Gold Standard*. Si consideramos las relaciones $R_1(M_1, E_1)$ y $R_2(M_2, E_2)$, donde M_i y E_i representan, respectivamente, el medicamento y el efecto adverso contenidos en la relación i, existen casos para los que E_1 y E_2 son dos formas diferentes de definir el mismo efecto adverso, es

decir, contienen palabras muy similares, aunque no son exactamente iguales. Por ejemplo, uno de los efectos adversos del medicamento *"itraconazole"* se define como *"vanishing bile duct"*, así como *"vanishing bile duct syndrome"*, es decir, se contabilizan como dos relaciones pero únicamente difieren en una palabra y representan el mismo efecto adverso. Si observamos el *Gold Standard* nos damos cuenta de que la relación correcta sería entre el medicamento *"itraconazole"* y el efecto adverso *"vanishing bile duct syndrome"*.

La consecuencia de que se produzca esta situación, en términos de la evaluación del sistema, es que la precisión (y por tanto la Medida-F) disminuye debido al número de relaciones extraídas por el sistema, que aunque no existan en el *Gold Standard*, su efecto adverso es equivalente al de una relación que sí se encuentra en él (con el mismo medicamento). Para solucionar estos casos, se consideran aquéllas relaciones $R_1(M_1, E_1)$ y $R_2(M_2, E_2)$, obtenidas por el sistema, para las cuales $E_1 \equiv E_2$. Puesto que E_1 y E_2 son términos compuestos por una o más palabras, basamos esta equivalencia en la cercanía entre ambos números, según la medida de similitud de Jaccard, la cual, dados dos conjuntos, se expresa mediante la siguiente fórmula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \qquad (6)$$

donde, en nuestro caso A y B son las dos entidades (efectos adversos) que queremos comparar. Tras una serie de pruebas, se ha optado por establecer un valor mínimo de $J = 0{,}6$ para considerar que dos entidades se refieren al mismo efecto adverso. Este valor es lo suficientemente elevado como para que las equivalencias que introduce sean correctas y la precisión no disminuya. Una vez que se analiza la distancia entre efectos adverso a un mismo medicamento, se unen en una sola relación aquéllos efectos adversos con similitud superior al umbral. La Figura 1 muestra el resultado de aplicar el algoritmo de similitud sobre un subconjunto de efectos adversos provocados por el mismo medicamento ("Dalteparin").

Como podemos observar, los efectos adversos que se han unido en una sola entidad utilizan diferentes definiciones para representar a dicha entidad. En este caso, se considera que una relación del *Gold Standard* ha sido encontrada por nuestro sistema, para un medicamento concreto, si existe un efecto adverso dentro del conjunto de efectos adversos similares para ese medicamento, según la distancia Jaccard, cuya definición coincide exactamente con el efecto adverso de la relación del *Gold Standard*. Cada relación encontrada por el sistema, se considera a efectos de evaluación como una sola instancia.

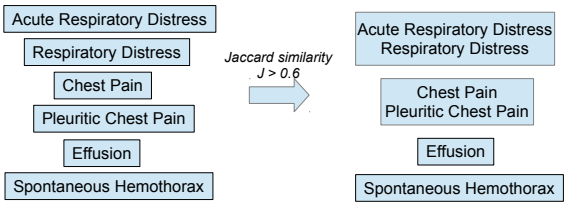

Figura 1: Aplicación de la medida de similitud Jaccard sobre los efectos adversos del medicamento *"Dalteparin"*.

La Tabla 3 muestra los resultados una vez aplicada la similitud de Jaccard entre efectos adversos, tanto a las relaciones obtenidas anteriormente como a las obtenidas por el *baseline*.

Sistema	P	C	F
BaseJac	27,11	**100,00**	42,65
PropJac	**45,46**	59,67	**51,60**

Tabla 3: Resultados tras aplicar la similitud de Jaccard sobre los efectos adversos, en función de la Medida-F (F), Precisión (P) y Cobertura (C). Comparación entre nuestro algoritmo (**PropJac**) y el *baseline* (**BaseJac**). Los campos en negrita indican el mayor valor de cada medida.

La tabla muestra un aumento de la Precisión y la Medida-F, gracias a la reducción de relaciones propuestas por el algoritmo, mientras que la Cobertura se mantiene constante, es decir, se sigue encontrando el mismo número de relaciones correctas que en los resultados iniciales.

5. *Inducción de conocimiento*

Uno de los aspectos más importantes de un sistema que busca relaciones dentro de textos, ya sea entre medicamentos y efectos adversos, como es el caso, o entre cualquier otro tipo de entidades, es su capacidad de descubrir nuevas relaciones que no se conociesen anteriormente, es decir inducir conocimiento nuevo. En el presente trabajo, es importante conocer si el sistema propuesto sería capaz de encontrar relaciones "nuevas", que no estén directamente basadas en evidencias que se puedan encontrar en los textos de partida. Dichas evidencias serían resúmenes concretos en los que aparezcan frases relacionando directamente medicamentos con efectos adver-

sos. En esta sección detallaremos los experimentos llevados a cabo para comprobar si nuestro sistema es capaz de realizar esta inducción de conocimiento.

El modelo de representación de conocimiento descrito en la Sección 3 nos permite obtener un conjunto de pares de entidades, relacionadas entre sí con un determinado peso. A partir de estos datos es sencillo construir un grafo en el que cada nodo sea una entidad (medicamento o efecto adverso), y un enlace entre dos nodos represente la significatividad estadística de coaparición de dichas entidades. El valor del enlace será el del peso almacenado para el par de entidades. Visualizando el modelo de representación de conocimiento como un grafo, los resultados obtenidos en la Sección 4.2 vendrían dados por las relaciones extraídas al recorrer sólo los enlaces directos entre medicamentos y efectos adversos del grafo, es decir, aquellos caminos de distancia $d = 1$ entre un medicamento y un efecto adverso. Sin embargo, es posible que si recorremos el grafo con mayor profundidad, encontremos nuevas relaciones. Sin embargo, es indispensable establecer condiciones para evitar que el número de relaciones propuestas aumente demasiado, comprometiendo la precisión del sistema (aunque se aumente su cobertura), y por tanto, la Medida-F. De acuerdo a esto último, en este caso se extraen las relaciones que se extraían anteriormente, y además, se añaden aquéllas en las que el medicamento y el efecto adverso se encuentran a dos pasos (enlaces) de distancia dentro del grafo. Además, consideramos que la entidad intermedia en dicho camino ha de ser otro efecto adverso: $M \Rightarrow E_1 \Rightarrow E_2$. Esta restricción parece lógica, basándonos en la hipótesis de que si E_1 y E_2 coaparecen con frecuencia, hay una probabilidad alta de que si M provoca E_1, pueda provocar también E_2.

Una vez que se generan nuevas relaciones gracias al grafo de coaparición, queremos saber si alguna de esas nuevas relaciones podría representar un caso de inducción de conocimiento. Como la única manera de saber si una relación es correcta sin recurrir a expertos del dominio es el *Gold Standard*, el proceso que se sigue para determinar si existe inducción de conocimiento es el siguiente:

1. Se extraen las nuevas relaciones, recorriendo el grafo con $d = 2$.
2. Se selecciona una relación concreta $R(M, E)$, que sea correcta según el *Gold Standard*, y se buscan aquéllos resúmenes que apoyen dicha relación. Es decir, se seleccionan los resúmenes en los que coaparezcan directamente el medicamento M y el efecto adverso E.
3. Se eliminan los resúmenes que apoyan la relación, y se vuelve a construir el grafo de coaparición.
4. Se vuelven a extraer las relaciones, con $d = 2$, del nuevo grafo de coaparición.
5. Si se vuelve a encontrar la relación $R(M, E)$, el conocimiento representado por dicha relación se ha inducido, sin que haya un resumen específico que apoye la existencia de dicha relación.

Estos pasos se han seguido para analizar varias relaciones positivas. Algunas de ellas se volvían a encontrar al final del proceso (se inducía el conocimiento), mientras que otras no. En este punto, nos interesa conocer si el peso de las relaciones juega un papel importante a la hora de inducir el conocimiento. Nuestra hipótesis se basa en que si el peso normalizado de una relación (el peso directo en el grafo, dividido entre la suma de los pesos de todos los enlaces que parten del nodo) es alto, dicha relación será más fuerte y por tanto el conocimiento que representa será más fácil de inducir. La forma de obtener el peso normalizado de un enlace entre el nodo i y el j, $P(i, j)$, se muestra en la fórmula 7:

$$P(i,j) = \frac{D(i,j)}{\sum_{k=1}^{O(i)} D(i,k)}, \qquad (7)$$

donde $D(i, j)$ es el peso directo (sin normalizar) entre los nodos, y $O(i)$ es el número de enlaces que parten del nodo i (su *"Outdegree"*).

Por ejemplo, a partir del medicamento *"methotrexate"* encontramos tres relaciones diferentes, con distancia $d = 2$, que se vuelven a encontrar al final del proceso anterior (su conocimiento se induce):

- $methotrexate \Rightarrow toxicity(0{,}05) \Rightarrow renal\ toxicity(0{,}48)$
- $methotrexate \Rightarrow sarcoma(0{,}05) \Rightarrow nodules(0{,}03)$
- $methotrexate \Rightarrow arthritis(0{,}31) \Rightarrow nephropathy(0{,}21)$

Los números situados a la derecha de los efectos adversos nos indican el peso normalizado de la relación entre la entidad anterior y el efecto adverso. De las tres relaciones mostradas, únicamente la última relación presenta ambos pesos

normalizados relativamente elevados (uno representa el 30 % de los pesos y el otro el 20 %, aproximadamente). Por tanto, podemos establecer como condición, que si se cumplen las restricciones $P(M_1, E_1) > 0{,}3$ y $P(E_1, E_2) > 0{,}2$ entonces existe $R(M_1, E_2)$ (el sistema considera la relación como correcta).

Una vez determinados estos umbrales, se vuelven a considerar todos los resúmenes para construir el grafo de coaparición. Esta configuración del sistema se ha denominado "Normalizada 1" o "N1".

Nos interesa, igualmente, realizar una prueba en el que se restrinja al máximo uno de los dos pesos, en este caso el peso normalizado entre E_1 y E_2, obligando a que el peso de dicha relación sea igual a 1. La interpretación de este peso sería una relación directa entre dos efectos adversos, en la que E_1 únicamente está conectado con E_2, y por tanto podemos suponer que la ocurrencia del efecto E_1 provoca en todos los casos que ocurra también el efecto E_2. Esta configuración del sistema se denomina "Normalizada 2" o "N2".

Por último, para comprobar el comportamiento base del grafo de coaparición con $d = 2$, consideramos la configuración del sistema "Normalizada 0" o "N0". En esta configuración no se restringen los pesos normalizados, sino que se generan todas las posibles relaciones $R(M, E_2)$ y se añaden a las obtenidas en los resultados iniciales.

La Tabla 4 contiene los resultados de todas las configuraciones del sistema, en términos de Precisión, Cobertura y Medida-F. Se ha añadido el número de relaciones correctas encontradas (*True Positives*) para ilustrar en términos absolutos el comportamiento de los sistemas.

Se observa que la configuración **N0** (sin restricciones en los pesos) obtiene el mayor valor de cobertura posible, encontrando más relaciones que ninguna otra. Sin embargo, el número de relaciones totales que obtiene es demasiado elevado, por lo que su precisión y Medida-F finales son muy pequeñas. La configuración **N1** consigue un compromiso entre precisión y cobertura que provoca que la Medida-F se mantenga similar a la conseguida por la configuración inicial. El aspecto importante de esta configuración es la inducción de conocimiento que se produce, tal y como se ha mostrado anteriormente. Finalmente, la configuración **N2** presenta el mejor valor de la Medida-F, aunque en este caso no se han encontrado casos en los que se produzca inducción de conocimiento.

En la tabla también se incluyen como referencia los valores de precisión, cobertura y Medida-F obtenidos por otros sistemas: un sistema (Gurulingappa, Mateen-Rajput, y Toldo, 2012) basado en máquinas de soporte vectorial (**JSRE**), otro sistema supervisado, aunque con tamaños pequeños del conjunto de entrenamiento, que utiliza bases de conocimiento (Kang et al., 2014), identificado en la tabla como **KB**, y un sistema basado en correspondencia de patrones (Eltyeb y Salim, 2015), también supervisado e identificado en la tabla como **PB**. Aunque los resultados ofrecidos por nuestro sistema quedan lejos de los obtenidos por dichos sistemas supervisados, estas técnicas requieren de unos recursos determinados para la fase de entrenamiento que el sistema propuesto en este trabajo no necesita.

Sistema	P	C	F	TP
Ini	**45,46**	59,67	51,60	3042
N0	10,45	**80,25**	18,48	**4091(*)**
N1	42,26	61,46	50,08	3133(*)
N2	45,12	60,34	**51,63**	3076
JSRE	86,00	89,00	87,00	—
KB	91,80	86,10	88,80	—
PB	93,60	72,80	81,70	—

Tabla 4: Resultados finales del sistema y comparación con otros sistemas. Se comparan los valores de cobertura (C), precisión (P) y Medida-F (F), expresados en porcentaje, así como el total de relaciones correctas encontradas (TP), sobre las cuatro configuraciones de nuestro sistema: inicial (**Ini**), normalizada 0 (**N0**), normalizada 1 (**N1**) y normalizada 2 (**N2**). Los campos en negrita indican el mayor valor de cada medida; el asterisco indica que se ha producido inducción de conocimiento. Se comparan nuestros resultados con otros sistemas, en este caso supervisados (ver texto).

6. Conclusiones y Trabajo Futuro

La técnica de extracción de relaciones descrita en este trabajo ofrece mejoras significativas en relación al *baseline* propuesto, lo cual nos indica que nuestro sistema discrimina correctamente aquéllas coapariciones de medicamentos y efectos adversos susceptibles de convertirse en una relación. La eficacia del sistema tiene un elevado margen de mejora, como se puede observar en la cobertura potencial que se podría alcanzar utilizando el grafo de coaparición (Tabla 4). El análisis de las relaciones que se podrían extraer a través de exploraciones más profundas del grafo (con valores de d mayores que 2) es una de las

cuestiones más inmediatas a analizar. También se analizarán patrones que permitan diferenciar las relaciones que nos interesan, de aquéllas que representan medicamentos aplicados como tratamiento a enfermedades específicas. En lo relativo al análisis de distancias, se explorarán otras técnicas orientadas a la extracción de equivalencias entre enfermedades, como por ejemplo el uso de variaciones léxicas propuestas por bases de datos médicas como SNOMED (Donelly, 2006).

Es importante destacar que las relaciones que se extraen en este trabajo se cotejan con un *Gold Standard* determinado, extraído a partir de un corpus que obviamente, no contiene todas las posibles relaciones existentes entre medicamentos y efectos adversos. Por tanto, es posible que exista conocimiento inducido que responde a relaciones correctas, pero que no se pueden evaluar como tales por falta de documentos médicos que las apoyen. En este sentido, sería útil realizar búsquedas en diversas fuentes, como bases de datos que proporcionen artículos médicos, para comprobar si existen evidencias de que una relación encontrada en el grafo pero no presente en el corpus ADE, puede ser igualmente correcta.

Bibliografía

Aramaki, E., Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, y K. Ohe. 2010. Extraction of adverse drug effects from clinical records. *Studies in health technology and informatics*, 160(Pt 1):739–743.

Donelly, K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279–290.

Edwards, I. R. y J. K. Aronson. 2000. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet*, 356(9237):1255 – 1259.

Eltyeb, S. y N. Salim. 2015. Pattern-based system to detect the adverse drug effect sentences in medical case reports. *Journal of Theoretical and Applied Information Technology*, 71(1):137–143.

Gurulingappa, H., J. Fluck, M. Hofmann-Apitius, y L. Toldo. 2011. Identification of adverse drug event assertive sentences in medical case reports. En *Proceedings of First international workshop on knowledge discovery and health care management (KD-HCM)*, páginas 16–27, Athens.

Gurulingappa, H., A. Mateen-Rajput, y L. Toldo. 2012. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Gurulingappa, H., A. Mateen Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, y L. Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892.

Kandula, S. y Q. Zeng-Treitler. 2010. Exploring relations among semantic groups: a comparison of concept co-occurrence in biomedical sources. *Studies in health technology and informatics*, 160(2):995–999.

Kang, N., E.M. van Mulligen, B. Singh, C. Bui, Z. Afzal, y J. Kors. 2014. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):64.

Lindberg, D., B. Humphreys, y A. McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

Pyysalo, S., A. Airola, J. Heimonen, J. Bjorne, F. Ginter, y T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6+.

Segura-Bedmar, I., S. de la Peña González, y P. Martínez. 2014. Extracting drug indications and adverse drug reactions from spanish health social media. En *Proceedings of BioNLP 2014*, páginas 98–106, Baltimore, Maryland, June. Association for Computational Linguistics.

Wang, X., G. Hripcsak, M. Markatou, y C. Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *JAMIA*, 16(3):328–337.