Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 75-82 recibido 05-05-12 revisado 01-06-12 aceptado 05-06-12

Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en español*

Analyzing the Use of Shallow Semantic Knowledge with Similarity Methods for Mapping Nursing Information in Spanish

Jorge Cruanes
Dep. Leng. y Sist. Inf.
Universidad de Alicante
03690, Alicante(Spain)
jcruanes@dlsi.ua.es

M. Teresa Romá-Ferri
Dep. Enfermería
Universidad de Alicante
03690, Alicante(Spain)
mtr.ferri@ua.es

Elena Lloret Pastor
Dep. Leng. y Sist. Inf.
Universidad de Alicante
03690, Alicante(Spain)
elloret@dlsi.ua.es

Resumen: Uno de los problemas actuales en el dominio de la salud es reutilizar y compartir la información clínica entre profesionales, ya que ésta se encuentra escrita usando terminologías específicas. Una posible solución es usar un recurso de conocimiento común sobre el que mapear la información existente. Nuestro objetivo es comprobar si la adición de conocimiento semántico superficial puede mejorar los mapeados establecidos. Para ello experimentamos con un conjunto de etiquetas de NANDA-I y con un conjunto de descripciones de SNOMED-CT en castellano. Los resultados obtenidos en los experimentos muestran que la inclusión de conocimiento semántico superficial mejora significativamente el mapeado léxico entre los dos recursos estudiados.
Palabras clave: PLN, similitud léxica, salud, enfermería, NANDA-I, SNOMED-CT

Abstract: In health domain, one of the current problems is the reusing and the sharing the clinical information between professionals, due to this information is written using specific terminologies. One possible solution is to use a common knowledge resource for mapping the existing information. In this paper, our aim is to analyze if the use of lexical similarity algorithms enriched with shallow semantic knowledge can improve these mappings. In order to achieve this, we experiment with a set of NANDA-I labels and a set of SNOMED-CT descriptions in Spanish. The results obtained show that the addition of shallow semantic knowledge significantly improves the lexical mapping between both studied resources.
Keywords: NLP, lexical similarity, health, nursing, NANDA-I, SNOMED-CT

1. Introducción

El reto actual en el dominio de la salud es contar con sistemas que permitan una rápida y eficiente manera de compartir información. Esta información en el caso específico de enfermería es fundamental, ya que suelen reflejar tanto los problemas del paciente o los diagnósticos de enfermería como las intervenciones realizadas y los resultados obtenidos.

Con el objetivo de conseguir esta eficiencia en la comunicación entre los profesionales de la salud, muchos países han comenzado a utilizar la Historia Clínica Electrónica (HCE). Estos sistemas permiten a los profesionales incluir texto estructurado y texto libre en los apartados para narración, empleando sus propios términos. Esto conlleva el empleo de diferentes palabras o etiquetas para referirse a un mismo concepto, generando una heterogeneidad en la información que dificulta compartirla y reutilizarla (Romá-Ferri y Palomar, 2008; Zwaanswijk et al., 2011).

Una posible solución pasa por la existencia de un recurso de conocimiento que permita la interconexión de las terminologías existentes. Para conseguir mapear la información escrita en lenguaje natural sobre estos recursos de

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria d'Educació de la Generalitat Valenciana (proyectos PROMETEO/2009/119, ACOMP/2010/286 y ACOMP/2011/001). Damos las gracias a los revisores por los comentarios constructivos, que nos han permitido mejorar este trabajo.

conocimiento es necesario el uso de métodos automáticos, debido a la gran cantidad de información existente así como a las constantes actualizaciones de las terminologías. En la actualidad, una de las soluciones adoptadas es la aplicación de métodos de similitud léxica (Meizoso, Allones y Taboada, 2011), ya que éstos son independientes del idioma y no requieren de corpus previamente etiquetados. Estas técnicas determinan cuán relacionadas están entre sí dos expresiones en base a su similitud de escritura. Sin embargo, están limitadas, pues no son capaces de discernir la semántica.

Tratando de ayudar a superar este problema, en este trabajo proponemos una aproximación de mapeado de información en lenguaje natural en el dominio de enfermería en castellano, basada en el uso de métodos de similitud léxica junto con conocimiento semántico superficial. El uso de semántica superficial pretende mejorar la evaluación de la similitud así como aportar una mayor eficacia en el mapeado.

Para verificar la eficacia de nuestra propuesta usaremos como origen la terminología NANDA-I (*North American Nursing Diagnosis Association International*) (NANDA-I, 2010) y como destino la terminología SNOMED-CT (*Systematized Nomenclature of Medicine-Clinical Terms*) (International Health Terminology Standards Development Organisation -IHTSDO-, 2010). NANDA-I consiste en una lista de diagnósticos de enfermería expresados en lenguaje natural y SNOMED-CT es una terminología clínica integral, formalizada y que permite la interoperabilidad semántica entre distintos sistemas, dando soporte a la diversidad lingüística. La terminología SNOMED-CT es la terminología de referencia que ha sido seleccionada para la HCE del Sistema Nacional de Salud español[1].

Este artículo está organizado en cinco secciones diferentes, comenzando con esta introducción. En la sección 2 se describe el estado de la cuestión. Las secciones 3, 4 y 5 describen, respectivamente, los materiales usados en las pruebas, nuestra propuesta y los experimentos que fueron realizados. La sección 6 muestra las conclusiones y se proponen mejoras para futuras investigaciones.

2. Estado de la cuestión

Para solucionar el problema de mapear diferentes terminologías y texto en lenguaje natural entre sistemas, una de las estrategias es hacer uso de técnicas de similitud léxica. Aunque no es la única, y se han empleado otras estrategias basadas en jerarquías de ontologías (Meizoso, Allones y Taboada, 2011) o métodos estadísticos (Nyström et al., 2010). Sin embargo, en el dominio de enfermería en castellano estas alternativas no son una opción inicialmente. En este ámbito no se cuenta con terminologías formalizadas que especifiquen explícitamente el significado de sus términos, ni con corpus etiquetados.

En el mapeado de información a SNOMED-CT mediante uso de técnicas de similitud léxica en lengua inglesa podemos destacar los trabajos de Patrick y Budd (2006), Wang et al. (2006), Patrick, Wang y Budd (2007) y Stenzhorn et al. (2009). Para el castellano podemos destacar el trabajo de Farfán Sedano et al. (2009). En estos trabajos la terminología origen para el mapeo es diversa. En unos casos se optó por las expresiones incluidas en las notas clínicas (Patrick y Budd, 2006; Patrick, Wang y Budd, 2007) o bien por la información sobre medicamentos incluida en una base de datos de farmacia hospitalaria (Farfán Sedano et al., 2009).

Una de las características comunes en los estudios sobre similitud léxica es la de realizar un preprocesamiento (normalización) a los textos a comparar. Esta preparación está ligada a la necesidad de, por ejemplo, reducir errores que se producen al comparar expresiones en mayúsculas con expresiones en minúsculas. Uno de los preprocesamientos más comunes es el de eliminar las palabras conocidas como 'stopwords' y signos de puntuación (Wang et al., 2006).

En cuanto a las técnicas de similitud más usadas son la similitud léxica exacta (Patrick y Budd, 2006; Wang et al., 2006; Patrick, Wang y Budd, 2007; Farfán Sedano et al., 2009) y el algoritmo del Coseno (Stenzhorn et al., 2009). En algunos casos las cadenas comparadas son expandidas para cubrir ciertas variaciones léxicas, como por ejemplo sustituir las abreviaturas por sus expresiones completas (Wang et al., 2006). Otros métodos hacen uso de un lexicón para mejorar el

[1]Sistema Nacional de Salud español. Referencia Web: http://www.msps.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/snomedHCD.htm. Último acceso 5/5/2012.

rendimiento en la búsqueda. En los trabajos de Patrick y Budd (2006) y de Patrick, Wang y Budd (2007) este lexicón es un índice de todas las palabras existentes en SNOMED-CT, asociadas con los identificadores de concepto en los que aparecen. La búsqueda consiste en encontrar el identificador de SNOMED-CT que contenga la mayor subcadena coincidente.

En algunos trabajos existe una etapa de postprocesamiento donde vuelven a usarse técnicas de similitud léxica. Esta etapa persigue mejorar los mapeados, bien combinando elementos entre sí para lograr un término más general en SNOMED-CT que los englobe (Patrick y Budd, 2006; Patrick, Wang y Budd, 2007), o bien realizando comparaciones de subcadenas para aquellas expresiones para las que no se ha encontrado un equivalente satisfactorio (Wang et al., 2006).

Tras esta revisión podemos observar que las técnicas de similitud léxica han sido ampliamente usadas en los mapeados de terminologías en el dominio de la salud. Sin embargo, los trabajos estudiados presentan algunas carencias, como la falta de consideración semántica de las palabras. Por ejemplo, las expresiones "alteración nutricional por deficiencia" y "alteración nutricional por exceso", aunque son léxicamente parecidas, son opuestas en significado. Otra carencia existente es por la omisión de stopwords, algunas de las cuales pueden llegar a cambiar drásticamente la semántica y, por tanto, el resultado de la similitud. Por ejemplo, en una comparación léxica de las expresiones "alimento con gluten" y "alimento sin gluten", si no se tienen en cuenta las stopwords se estaría comparando la misma expresión: "alimento gluten". Es por ello, que la evaluación a nivel léxico no es suficiente, y es necesario aportar, al menos, una capa de semántica superficial.

3. Materiales

Para poder analizar el uso de conocimiento semántico superficial sobre métodos de similitud léxica, hemos creado una serie de materiales de trabajo. Con estos materiales se ha diseñado una batería de pruebas para comprobar la eficiencia de nuestra aproximación.

El primer material de trabajo está basado en la terminología NANDA-I en castellano. Esta terminología identifica una serie de diagnósticos de enfermería estandarizados para el cuidado de pacientes. Cada diagnóstico es una etiqueta que describe una situación y cuenta con un código propio que la identifica. El material de trabajo se creó a partir de la recopilación de las etiquetas de NANDA-I en castellano de las versiones 2001-2002, 2005-2006, 2007-2008 y 2009-2011. Con ello se recopilaron todas las descripciones y variantes léxicas existentes (variaciones de términos y signos de puntuación). La recopilación aportó un total de 728 etiquetas, de las cuales hemos utilizado seis para evaluar manualmente los resultados obtenidos en la experimentación.

El segundo material de trabajo consiste en un subconjunto de SNOMED-CT en castellano. SNOMED-CT es una terminología clínica, estructurada jerárquicamente, que proporciona contenido para informes y documentación clínica. En SNOMED-CT, cada concepto tiene un código único, una descripción completa (llamada *full*), una descripción preferente (*preferred*) y sus posibles sinónimos (*synonyms*). Cada concepto tiene una única representación terminológica (*full*), aunque incluye las diferentes etiquetas usadas en el dominio de salud para mencionar a dicho concepto (*synonyms*), incluso manteniendo aquellas que ya no están en uso. Por ejemplo, el concepto 85623003 tiene como descripción completa 'Termorregulación ineficaz (hallazgo)' y como descripción preferente tiene la etiqueta 'Termorregulación ineficaz'. Para nuestra experimentación se ha seleccionado un subconjunto de SNOMED-CT de su versión española del 30/04/2011, extrayéndolo a partir de los términos clave de los seis diagnósticos NANDA-I seleccionados. Este subconjunto está compuesto por 36 descripciones, que han sido evaluadas manualmente para establecer su equivalencia con las etiquetas NANDA-I. Tras la evaluación manual se estableció que solo nueve eran descripciones equivalentes a las etiquetas de NANDA-I.

Para las funciones de conocimiento semántico se crearon 3 recursos complementarios y con conocimiento ajustado a este dominio. Por una parte, se creó una bolsa con 68 antónimos y otra bolsa con 34 sinónimos; en ambos casos se incluyeron tanto sustantivos como adjetivos. Por otra parte, se añadió una 'expresión existencial'. Denominamos 'expresiones existenciales' a aquellas expresiones condicionales que modifican la semántica de toda la frase. No usamos recursos semánticos

de dominio abierto, como por ejemplo EuroWordNet, ya que, tras un estudio inicial, se detectó que no se ajustaban a las necesidades del dominio en estudio. Por ejemplo, las palabras 'exceso' y 'defecto', que en el dominio de enfermería son consideradas como antónimos, no lo son en ninguno de los recursos estudiados.

Finalmente, la experimentación se realizó usando 15 algoritmos de similitud léxica, mediante la implementación proporcionada por la librería de código abierto Java SimMetrics (versión 1.6.2) (Chapman, 2006). Los algoritmos usados han sido: (i) Coseno, (ii) Levenshtein, (iii) Similitud de Dice, (iv) Distancia Euclídea, (v) Similitud de Jaccard, (vi) Distancia Jaro-Winkler, (vii) Coeficiente de Matching, (viii) Needleman Wunch, (ix) Smith Waterman, (x) Coeficiente de Superposición, (xi) Monge Elkan, (xii) Distancia de Bloque, (xiii) Desviación de Distancia de Chapman, (xiv) Q Grams Distance y (xv) Soundex.

4. Método

En esta sección mostramos nuestra propuesta para el mapeado entre etiquetas de NANDA-I y SNOMED-CT para el idioma castellano. Para ello proponemos el uso de tres aproximaciones, complementarias al análisis de similitud léxica, para determinar la similitud entre etiquetas: (i) detección de antonimia, (ii) detección de sinonimia y (iii) detección de *'expresiones existenciales'*. Estas aproximaciones semánticas ayudan a establecer un grado de similitud entre dos etiquetas comparadas sin necesidad de entrenamiento ni corpus etiquetados. Las tres aproximaciones propuestas se han diseñado independientemente, de forma que pueden ser aplicadas de forma individual o combinándose entre ellas. El proceso se detalla en el Algoritmo 1.

Según se describe en el Algoritmo 1, el conocimiento semántico superficial se aplica en las funciones 'areRelated', 'areAntonyms' y 'getSynonyms', mientras que la función 'getSimilarity' devuelve el resultado de la comparación léxica de las etiquetas. El orden aplicado se corresponde a la mayor eficiencia de su ejecución en caso de que se haga uso de todas las funciones de conocimiento semántico superficial definidas. A continuación se detalla el funcionamiento de cada una de las funciones empleadas.

La función 'areRelated' se encarga de comprobar si hay expresiones en las etiquetas que establezcan su 'no equivalencia' semántica. Por ejemplo, si una de las etiquetas es "Riesgo de desequilibrio nutricional" y otra es "Desequilibrio nutricional", se establecen como 'no equivalentes', puesto que la primera expresa una probabilidad, mientras que la segunda expresa un hecho. En el caso de dos etiquetas con una 'expresión existencial' se consideran como posibles equivalentes cuando en ambas está presente dicha 'expresión existencial' o un sinónimo.

Algoritmo 1 Obtención de la similitud entre etiquetas, usando comparación léxica con conocimiento semántico superficial.

```
if areRelated(label1, label2) then
  if areAntonyms(label1, label2) then
    return 0;
  else
    {Expansión de sinónimos}
    labelSynons1=getSynonyms(label1);
    labelSynons2=getSynonyms(label2);
    {Para todas las expansiones}
    for all syn1 in labelSyns1 do
      for all syn2 in labelSyns2 do
        {Comparación léxica}
        tmpScore = getSimilarity(syn1,syn2);
        {Guardamos el mejor resultado}
        if tmpScore > maxScore then
          maxScore = tmpScore;
        end if
      end for
    end for
    return maxScore;
  end if
  return 0;
end if
```

La función 'areAntonyms' se encarga de comprobar la antonimia de las etiquetas. Por ejemplo, si una etiqueta es "Aumento de la fiebre" mientras que la otra es "Disminución de la fiebre", con significado antónimos, son establecidas como 'no equivalentes'.

La función 'getSynonyms', a partir de una etiqueta origen, devuelve un conjunto de etiquetas sinónimas generadas por la expansión de sus sinónimos. Por ejemplo, expandiendo la etiqueta "Aumento de la fiebre" obtendríamos como sinónima la nueva etiqueta "Incremento de la fiebre". Todas las expansiones de las etiquetas son usadas en la comparación léxica, buscando maximizar la

Prueba	Etiqueta NANDA-I	Etiqueta SNOMED-CT	Resultado ideal
I	1 etiqueta NANDA-I en uso	1 descripción completa y 1 preferente	Ambas son equivalentes a la etiqueta NANDA-I
II	1 etiqueta NANDA-I en desuso	1 descripción completa y 1 preferente	Ambas son equivalentes a la etiqueta NANDA-I
III	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente
IV	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente
V	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Sólo una descripción completa, una preferente y tres sinónimas son equivalentes a la etiqueta NANDA-I
VI	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente

Tabla 1: Descripciones de las pruebas diseñadas y resultado ideal esperado.

Prueba	Etiqueta NANDA-I	Etiqueta SNOMED-CT	Singularidades
I	Termorregulación ineficaz	Termorregulación ineficaz	Iguales léxicamente
III	Desequilibrio nutricional: ingesta superior a las necesidades	Riesgo de desequilibrio nutricional: ingesta superior a las necesidades	Aunque tienen alta similitud léxica, la etiqueta SNOMED-CT está escrita como probable y la de NANDA-I como hecho
V	Riesgo de desequilibrio nutricional: ingesta superior a las necesidades	Riesgo de desequilibrio nutricional: ingesta inferior a las necesidades	Aunque tienen alta similitud léxica, son opuestas en significado
VI	Riesgo de infección	Trastorno nutricional: potencial de exceso para los requerimientos corporales	Léxicamente muy diferentes. Fue diseñado como prueba de control.

Tabla 2: Ejemplos de etiquetas usadas en los experimentos junto con sus singularidades.

puntuación obtenida.

Finalmente, la función 'getSimilarity' se encarga de evaluar la similitud léxica entre dos etiquetas. Esta función devuelve un valor entre 0 y 1. Estos valores representan el grado de similitud entre dos expresiones, lo cual consideraremos como el resultado de la comparación. Los resultados con un valor de 1 indican que las dos etiquetas son completamente equivalentes.

4.1. Diseño de los experimentos

Con el fin de comprobar la validez del método propuesto se diseñó un conjunto de seis pruebas. Puesto que nuestro objetivo es analizar la inclusión de conocimiento semántico superficial, como valores de referencia en la evaluación usamos los resultados de la comparación de similitud léxica simple (valores base). Para cada prueba se seleccionaron manualmente una etiqueta NANDA-I y varias de SNOMED-CT (descripciones completas, preferentes y sinónimas).

El esquema de cada prueba se encuentra indicado en la Tabla 1, donde se detallan las etiquetas incorporadas así como el resultado ideal esperado. Para completar la descripción, en la Tabla 2, se muestran algunos ejemplos de las etiquetas empleadas en las pruebas y las singularidades a superar por nuestra propuesta. Por ejemplo, en la prueba V (Tabla 2), las palabras subrayadas mues-

tran que las etiquetas comparadas, aunque léxicamente sean muy parecidas, no pueden considerarse equivalentes.

Por una parte, tras la ejecución del algoritmo propuesto, se realizó un análisis manual de los resultados para determinar si las etiquetas comparadas debían catalogarse como 'equivalentes' o 'no equivalentes'. Por otra parte, se analizaron los resultados respecto a los valores de los umbrales. La hipótesis de partida fue que si, dado un cierto umbral, el resultado obtenido es igual o mayor a él, entonces el algoritmo establecía como 'equivalentes' las etiquetas comparadas.

Posteriormente, se relacionaron ambos análisis y las etiquetas comparadas que eran equivalentes reales se consideraron 'verdaderos positivos' cuando el resultado alcanzaba o superaba el umbral analizado, o bien 'falso negativos' cuando no llegaba al umbral. De igual forma, cuando dos etiquetas comparadas no son en realidad equivalentes, se consideraron como 'falso positivo' cuando el resultado alcanzaba o superaba el umbral y, en caso contrario, se consideraba 'verdadero negativo'. Por ejemplo, al aplicar el algoritmo para comparar las etiquetas "Riesgo de infección" e "Infección Potencial", manualmente catalogadas como equivalentes, se obtiene como resultado 0.68, si se analizaba el resultado utilizando el umbral de 0.65, entonces se consideraba un resultado 'verdadero positivo', al ser equivalentes las etiquetas comparadas. Sin embargo, si se empleaba como umbral 0.7 el resultado sería considerado como un 'falso negativo', pues no sería identificada la comparación como 'equivalente'.

Los valores de los umbrales se fijaron entre 0.3 y 0.85 con un incremento de 0.05 para el análisis de los resultados. Esto se estableció acorde a un estudio piloto, donde se comprobó que los valores de los umbrales inferiores a 0.3 producían gran cantidad de falsos positivos (ruidos para los potenciales usuarios) y los umbrales superiores a 0.85 eran excesivamente restrictivos, obteniendo verdaderos positivos para aquellos casos en los que la etiqueta comparada tenía una escritura casi idéntica (silencio o perdida de información significativa para los usuarios).

5. Resultados

Para determinar la eficacia del conocimiento semántico superficial aplicado sobre los 15 algoritmos estudiados, se analizaron los resultados en base a la cobertura y la precisión, en tres ejecuciones diferentes: (i) sin utilizar conocimiento semántico, (ii) usando conocimiento parcial de antonimia y 'expresiones existenciales' y (iii) usando conocimiento semántico superficial completo. Los algoritmos de similitud léxica que mejor respondieron a la detección de etiquetas 'equivalentes' o 'no equivalentes' se muestran en las Tablas 3 y 4. En cada caso, al nombre del algoritmo se le ha añadido la letra 'C' cuando se ha complementado con conocimiento semántico superficial completo, la letra 'P' cuando se ha empleado conocimiento parcial y sin especificar letra para la aplicación del algoritmo básico. En ambas tablas, cada columna representa los valores del umbral utilizado, y los resultados obtenidos son mostrados como porcentajes, resaltando los mejores resultados para cada umbral.

De entre los 15 algoritmos estudiados, los mejores valores de cobertura fueron aportados por Monge Elkan (ME) y Soundex (Tabla 3). Con el algoritmo de ME los mejores resultados se obtuvieron con una aportación parcial de conocimiento semántico, y se mejoró la cobertura entre un 6.2 % (umbrales entre 0.3 a 0.45) y un 12.5 % (umbrales superiores a 0.5). Los resultados base del algoritmo ME solo mejoraron, con conocimiento completo, los resultados en los umbrales 0.55 y 0.8 (un 12.5 % superior). Por contra, el algoritmo de Soundex obtuvo mejores resultados usando el conocimiento semántico completo; su mejora osciló entre 12.5 % para los umbrales 0.6 y 0.65 hasta el 7.1 % para los umbrales 0.8 y 0.85. Con conocimiento parcial, Soundex solo mejoró la cobertura un 6.3 % a partir del umbral 0.7.

Respecto a la precisión, los algoritmos que aportaron los mejores resultados base fueron Jaro y Levenshtein, mientras que los mejores complementados con conocimiento semántico fueron los algoritmos Levenshtein y Q Grams Distance (QGD) (Tabla 4)[2]. Los algoritmos Levenshtein y QGD se comportaron de forma idéntica con conocimiento parcial o completo respecto a los resultados base. El algoritmo Levenshtein aporta una mejora significativa situada por encima del 50 % para

[2]Se incluyen los resultados de QGD sin conocimiento semántico para facilitar la comparación de la mejora aportada por el conocimiento semántico P y C.

Algoritmos	**Umbrales**											
	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Monge Elkan	93.8	93.8	93.8	93.8	87.5	75.0	75.0	75.0	75.0	75.0	62.5	62.5
Monge Elkan P	**100**	**100**	**100**	**100**	**100**	87.5	87.5	87.5	**87.5**	**87.5**	**75.0**	**75.0**
Monge Elkan C	93.8	93.8	93.8	93.8	87.5	87.5	75.0	75.0	75.0	75.0	**75.0**	62.5
Soundex	**100**	**100**	**100**	**100**	**100**	**100**	87.5	87.5	62.5	62.5	50.0	50.0
Soundex P	**100**	**100**	**100**	**100**	**100**	**100**	87.5	87.5	68.8	68.8	56.3	56.3
Soundex C	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	71.4	71.4	57.1	57.1

Tabla 3: Resultados de cobertura aplicando algoritmos de similitud básicos, con conocimiento semántico superficial parcial (P) y completo (C).

Algoritmos	**Umbrales**											
	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Jaro	34.9	34.9	34.9	35.7	32.5	34.2	40.0	42.9	42.1	28.6	**100**	**100**
Levenshtein	27.6	27.6	30.8	36.4	44.4	33.3	22.2	50.0	33.3	**100**	**100**	**100**
Levenshtein P	88.9	88.9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Levenshtein C	88.9	88.9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Q Grams Distance	35.7	35.7	41.7	29.4	20.0	16.7	22.2	25.0	25.0	33.3	**100**	**100**
Q Grams Distance P	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Q Grams Distance C	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Tabla 4: Resultados de precisión aplicando algoritmos de similitud básicos, con conocimiento semántico superficial parcial (P) y completo (C).

los umbrales hasta 0.7 (mínima mejora del 50 % en el umbral 0.65 y máxima de 77.8 % en el umbral 0.6). El algoritmo QGD, aunque no obtuvo los mejores resultados base en precisión, la inclusión tanto de conocimiento parcial como completo mejoró los resultados entre un 58.3 % (umbral 0.4) y un 83.3 % (umbral 0.55), además de lograr una precisión del 100 % para los umbrales entre 0.3 y 0.75.

5.1. Discusión

La inclusión de conocimiento semántico superficial mejora significativamente la comparación léxica simple, llegando a mejorar hasta en un 12.5 % la cobertura y un 83.3 % la precisión, alcanzando porcentajes de cobertura del 87.5 % y precisión del 100 % para umbrales hasta el 0.75. Estos resultados suponen además una mejora significativa respecto al actual estado del arte, donde se reportan coberturas del 80 % y del 83 % para los trabajos de Wang et al. (2006) y Meizoso, Allones y Taboada (2011) respectivamente. En cuanto a precisión, estos dos mismos trabajos reportan precisiones del 50 % y del 95 % respectivamente. El resto de trabajos no aportan datos claros y completos respecto a su cobertura y precisión. Las principales mejoras aportadas por el método presentado frente a la comparación léxica exacta o el algoritmo del Coseno reside en que nuestra aproximación premia las subcadenas coincidentes y las similitudes parciales de cadenas, además de aportar comparación de sinonimia y restricciones semánticas.

Las penalizaciones impuestas por las restricciones de antonimia y 'expresiones existenciales' han prevenido emparejamientos incorrectos y, por lo tanto, mejorado significativamente los niveles de precisión respecto a los valores base (usar solo similitud léxica).

Sin embargo, estudiando detenidamente los resultados observamos que el uso de conocimiento semántico de sinónimos produce resultados dispares. Por un lado hace caer los resultados de cobertura con Monge Elkan, mientras en el caso del algoritmo de Soundex sirve para mejorarlos. En cuanto a los algoritmos de Levenshtein y Q Grams Distance no produce ningún cambio. Por otro lado, se detectó un efecto ruido en el experimento control diseñado. La sustitución de 'riesgo' por su sinónimo 'potencial' provocó un resultado superior en la compara-

ción y generó un falso positivo al superar los umbrales de análisis.

6. Conclusiones y trabajos futuros

Este artículo propone un método para el mapeado automático entre terminologías en el dominio de enfermería en castellano, combinándolo los métodos de similitud léxica clásicos con conocimiento semántico superficial en tres vías diferentes: sinónimos, antónimos y 'expresiones existenciales'.

Tras la experimentación hemos podido constatar que nuestra aproximación mejora los resultados de cobertura y de precisión respecto a métodos de similitud léxica, considerados como resultados base. Sin embargo, el uso de sinónimos en la expansión de etiquetas léxicamente muy diferentes puede producir ruido cuando comparten un término. Es por ello que esta aproximación debe ser mejorada para evitar mapeados erróneos.

Como trabajo futuro, para mejorar nuestra aproximación mediante el uso de sinónimos y prevenir falsos positivos, proponemos establecer un mínimo de términos en común entre dos etiquetas para poder ser comparadas. De esta forma, pretendemos evitar la comparación entre dos etiquetas no relacionadas. Este mínimo número de términos en común deberá ser establecido mediante experimentación, para establecer el margen más adecuado.

Bibliografía

Chapman, S. 2006. SimMetrics. Recuperado el 20 de Noviembre de 2011, desde http://sourceforge.net/projects/simmetrics.

Farfán Sedano, F. J., M. Terron Cuadrado, E. M. García Rebolledo, Y. Castellanos Clemente, P. Serrano Balazote y A. Gomez Delgado. 2009. Implementation of SNOMED CT to the medicines database of a general hospital. *Studies in Health Technology and Informatics*, 148:123–130.

International Health Terminology Standards Development Organisation -IHTSDO-. 2010. SNOMED Clinical Terms User Guide. Informe técnico, IHTSDO.

Meizoso, M., J. Allones y M. Taboada. 2011. Automated mapping of observation archetypes to SNOMED CT concepts. En *4th international conference on Interplay between natural and artificial computation, IWINAC 2011*, volumen 6686, páginas 550–561. Springer-Verlag, Berlin.

NANDA-I. 2010. *Diagnósticos Enfermeros: Definiciones y Clasificación, 2009-2011*. Elsevier, Barcelona.

Nyström, M., A. Vikström, G. H. Nilsson, H. Åhlfeldt y H. Örman. 2010. Enriching a primary health care version of ICD-10 using SNOMED CT mapping. *Journal of Biomedical Semantics*, 1:7. Doi:10.1186/2041-1480-1-7.

Patrick, J. y P. Budd. 2006. Automatic conversion of clinical notes into snomed ct at point of care. En J. Westbrook J. Callen G. Margelis y J. Warren, editores, *Proceedings of HIC2006 and HINZ2006*, páginas 209–213. Health Informatics Society of Australia (Aotea Centre, New Zealand).

Patrick, J., Y. Wang y P. Budd. 2007. An automated system for conversion of clinical notes into SNOMED clinical terminology. En *Proceedings of the fifth Australasian symposium on ACSW frontiers*, volumen 68, páginas 219–226. Australian Computer Society (Ballarat).

Romá-Ferri, M. T. y M. Palomar. 2008. Análisis de terminologías de salud para su utilización como ontologías computacionales en los sistemas de información clínicos. *Gaceta Sanitaria*, 22(5):421–433.

Stenzhorn, H., E. J. Pacheco, P. Nohama y S. Schulz. 2009. Automatic Mapping of Clinical Documentation to SNOMED CT. *Studies in health technology and informatics*, 150:228–232.

Wang Y., J. Patrick, G. Miller y J. O'Halloran. 2006. Linguistic mapping of Terminologies to SNOMED CT. En *Proceedings of Semantic Mining Conference on SNOMED*. Network of Excellence Semantic Mining (Copenhagen).

Zwaanswijk, M., R. A. Verheij, F. J. Wiesman y R. D. Friele. 2011. Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study. *BMC Health Services Research*, 11:256. Doi:10.1186/1472-6963-11-256.