Anotación Automática de Discapacidades en Documentos Científicos de Medicina*

Automatic Disabilities Labeling in Medical Scientific Documents

Carlos Valmaseda
NLP Group at UNED
28040 Madrid, Spain
carlosvalmaseda@gmail.com

Juan Martinez-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: Este artículo presenta una herramienta para la anotación de discapacidades en documentos científicos. La identificación de conceptos médicos presentes en documentos y, especialmente, la identificación de discapacidades, es una tarea compleja debido principalmente a la gran variedad de expresiones que pueden referirse a un mismo problema. Nuestra propuesta, implementa una herramienta de anotación automática similar a UMLS MetaMap Transfer (MMTx) para la extracción de conceptos biomédicos. Al igual que MetaMap, nuestro sistema genera diferentes variantes de una misma discapacidad con el objetivo de mejorar la cobertura, adaptadas al tipo de entidad considerado. Así, en la generación de variantes se han utilizado palabras de impedimento o limitación (delay, impairment, etc.), que combinadas con funciones corporales o cognitivas dan lugar a nuevas expresiones de discapacidad. Los primeros resultados del sistema sobre una pequeña colección de documentos científicos anotados manualmente indican el potencial del mismo.
Palabras clave: Anotación de conceptos biomédicos, dominio médico, extracción de información

Abstract: This paper presents a tool for the annotation of disabilities in scientific papers. The identification of medical concepts in documents and, especially, the identification of disabilities, is a complex task mainly due to the variety of expressions that can make reference to the same problem. Our proposal, implements an automatic annotation tool similar to UMLS MetaMap Transfer (MMTx) for extracting biomedical concepts. As MetaMap, our system generates different variants of the same disability aiming to improve coverage, and adapting them to the kind of entity considered. Thus, in the generation of variants we use "impairment words" (delay, impairment, etc.), which combined with physical or cognitive functions provide new expressions of disability. The first results of the system on a small collection of scientific papers manually annotated indicate the potential of the proposal.
Keywords: Biomedical concepts labeling, medical domain, information extraction

1 Introducción

El estudio de las relaciones existentes entre distintos elementos del dominio biomédico es fundamental para proseguir los avances en el área. Se está dedicando grandes esfuerzos a identificar algunas de estas relaciones, tales como las interacciones entre proteínas, las asociaciones genes-enfermedades o los efectos adversos a medicamentos. La forma de abordar estos problemas suele consistir en la identificación por parte de expertos de algunas de estas relaciones. Como se trata de una tarea muy lenta y costosa, en la actualidad se están aplicando técnicas de aprendizaje automático para identificar relaciones que puedan encontrarse en textos relacionados con el dominio biomédico. Tanto para abordar este problema, como para la búsqueda especializada de terminología relacionada con algún aspecto específico del dominio es fundamental la anotación de los conceptos correspondientes, ya sean estos enfermedades, genes, proteínas, etc. Existen otros problemas en los que se requiere la anotación, como por ejemplo la clasificación y agrupamiento de documentos, la búsqueda de respuestas, etc.

En este trabajo se aborda la anotación de

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

un tipo de concepto que no está recogido en trabajos previos, al menos no de forma específica. Se trata de la identificación de expresiones relativas a discapacidades. Aunque algunas discapacidades están incluidas entre los síntomas de algunas ontologías del dominio biomédico, sólo se trata de unos pocos casos y su identificación necesita que estén mencionados de una forma específica. En este trabajo se aborda este problema, que aunque comparte ciertas aspectos con la anotación de conceptos en el dominio biomédico, también presenta una problemática particular, ya que las referencias a discapacidades pueden expresarse con más libertad que las referencias a enfermedades, genes, proteínas, etc. Las referencias a discapacidades admiten todo tipo de variantes sintácticas, morfológicas, semánticas, etc. Por ejemplo, para una misma discapacidad, podrían encontrarse las siguientes variantes:

- No puedo mover la pierna izquierda
- Limitaciones motrices en las extremidades inferiores
- No le responde la pierna, etc.

Por ello, en este marco se hace más relevante la aplicación de técnicas de procesamiento del lenguaje natural (PLN). Nuestro objetivo, por tanto, es utilizar éstas técnicas para identificarlas y anotarlas en textos médicos con la mayor precisión posible.

Para seleccionar las expresiones que se corresponden con discapacidades en los textos de entrada, partimos de una adaptación del enfoque seguido por el sistema MetaMap (Aronson, 2001) para nuestro problema. Al tratarse de expresiones en un lenguaje mucho más libre, necesitamos recurrir a otras técnicas. En particular, consideramos la polaridad de los términos involucrados que combinados con expresiones de funciones corporales o cognitivas dan lugar a nuevas expresiones de discapacidad.

Como marco experimental nos hemos centrado en las discapacidades asociadas a enfermedades raras (ERs), por varias razones. Por una parte se trata de un problema de gran importancia dada la escasa información disponible sobre las discapacidades asociadas a ERs, dada la propia naturaleza de estas enfermedades. Por otra parte, Orphanet[1], la organización internacional sobre las ERs y medicamentos huérfanos ha creado una colección especializada de textos dedicados a los profesionales y proveedores de servicios sociales; la Orphanet Discapacidad Enciclopedia. Se centra en las discapacidades asociadas con una ER específica. Estas fichas de discapacidades proporcionan una breve visión general de los aspectos médicos de la enfermedad, validados por expertos médicos, e incluyen una descripción de las discapacidades que experimentan los pacientes. Esta información nos va a permitir evaluar los resultados de este trabajo.

Por último, en Orphanet están indizando las consecuencias funcionales de cada ER con el Orphanet Funtioning Thesaurus, una adaptación de la Clasificación Internacional del Funcionamiento, de la Discapacidad y de la Salud de Niños y Jóvenes versión (ICF-CY (Organization., 2007)), que incluye términos adicionales para describir trastornos cognitivos, del sueño, del temperamento y de la conducta.

El punto de partida, además de la colección de documentos en los que aparezcan términos de ese tipo, son las listas de expresiones relacionadas con discapacidades. Partimos del tesauro de discapacidades de Orphanet, y ampliamos su terminología con la lista de discapacidades asociadas a las ERs consideradas en el corpus.

En el resto de este artículo se presentan en primer lugar antecedentes de anotación de conceptos en textos médicos en la Sección 2. Después, en la Sección 3 se describe la forma en que se ha construido un corpus de ERs con discapacidades asociadas, que será utilizado para analizar y evaluar las propuestas que se hacen después para la anotación de discapacidades. En la Sección 4 se muestra el funcionamiento del sistema de anotación de conceptos médicos. Finalmente se presentan los resultados de la evaluación en la Sección 5 y las conclusiones en la Sección 6.

2 Antecedentes

En la actualidad, existen muy pocos analizadores adaptados al dominio médico en español. MetaMap Transfer (MMTx) (Aronson, 2001) es una aplicación que tiene dos funcionalidades destacadas, por un lado puede mapear textos médicos al metatesauro UMLS[2],

[1] http://www.orpha.net

[2] http://www.nlm.nih.gov/research/umls/

y por otro lado permite descubrir conceptos del metatesauro en documentos.

Este sistema aplica al texto de entrada un análisis léxico/sintáctico que conlleva los siguientes pasos: Tokenizador, etiquetado léxico y análisis sintáctico superficial e identificación de los núcleos de los sintagmas. Por cada frase extraída tras este análisis, se aplican cuatro pasos. El primero de ellos es la generación de variantes, en la que se buscan las variantes de todas las palabras de las frases. Después se identifican los candidatos en base a su correspondencia con el texto de entrada. Más tarde, se realiza una construcción de la correspondencia, en la que las candidatas encontradas en el paso anterior se combinan y evalúan para producir como resultado final las mejores correspondencias de las frases del texto. La evaluación que se realiza tanto en las correspondencias de candidatas como en las propuestas finales, es una combinación lineal de cuatro medidas de inspiración lingüística: centralidad, variación, cobertura y cohesión.

Debido a esta escasez de recursos en español, aparecieron trabajos como el de (Carrero et al., 2008) en el que trataron de adaptar MetaMap al español mediante la traducción de los textos al inglés, para luego aplicar la extracción de conceptos médicos mediante MetaMap. Posteriormente surgieron otros trabajos (Iglesias et al., 2008) que implementaban un sistema completo como MOSTAS. Este sistema de etiquetado morfo-semántico también realiza funciones de anonimización de textos y corrector ortográfico con el objetivo de permitir la identificación de términos clínicos mediante el uso de SNOMED CT. Castro et al. (2010) presentaron una propuesta para la anotación semántica de informes clínicos en español. Implementaron una herramienta similar a UMLS MetaMap Transfer (MMTx) para la identificación de conceptos médicos sobre la ontología en español SNOMED CT. En otro trabajo similar (Oronoz, de Ilarraza, y Torices, 2010; Oronoz et al., 2013) se ha desarrollado una herramienta de anotación que detecta entidades en el dominio biomédico. Sobre la base de Freeling, los autores enriquecen su léxico con términos biomédicos extraídos de diccionarios y ontologías. La evaluación fue realizada sobre medicamentos, sustancias y enfermedades. Vivaldi y Rodriguez (2010) crearon un sistema de extracción de términos que usa información semántica extraída de Wikipedia. El sistema fue probado sobre un corpus médico, y según los resultados, podría considerarse como un buen recurso para la extracción de términos médicos. Conrado et al. (2011) llevan a cabo una extracción automática de términos médicos, usando sintagmas nominales previamente reconocidos en textos médicos en español. Los autores, haciendo uso de SNOMED CT, demuestran que es posible extraer términos médicos usando sintagmas nominales específicos.

3 Creación de un corpus anotado

Inicialmente, para la construcción del corpus se han considerado las siguientes enfermedades raras: Síndrome de Angelman, Síndrome de Cockayne, Epidermolisis bullosa distrófica, Síndrome X Fragil, Enfermedad de Norrie, Síndrome de Pendred. En la actualidad Orphanet tiene un conjunto de enfermedades para las que una serie de expertos han asociado sus discapacidades. De esta forma, las enfermedades que hemos seleccionado para este trabajo han sido tomadas de dicho conjunto.

Como ejemplo, las discapacidades asociadas (inglés) al Síndrome de Angelman son las siguientes:

- very low learning ability
- difficulty to mimic
- difficulty to memorize the gestures
- almost non-existent language
- slow execution of the instructions
- high fatigue
- attention disorders
- concentration disorders
- can not be completely autonomous

Este corpus de referencia fue anotado por un grupo de 3 voluntarios en el que cada uno etiquetó las discapacidades encontradas en varios artículos científicos en español. Después del proceso de etiquetado, tan solo se consideraron las discapacidades que habían sido detectadas por al menos dos personas independientemente y para las que había habido acuerdo sobre la anotación. El acuerdo entre anotadores fue medido mediante el valor kappa de Fleiss obteniendo un 0.68. El corpus definitivo está compuesto de 15 artículos científicos completos.

4 Sistema de Anotación de Discapacidades

El sistema, que utiliza recursos externos para realizar algunas tareas de procesamiento del lenguaje, comienza con un procesamiento del metatesauro en el que se generan las variantes de las discapacidades que contiene. Después, dado un documento, identifica los sintagmas nominales y genera sus variantes. Es decir, se generan variantes tanto de las discapacidades como de los candidatos en el documento. Por tanto es posible configurar los niveles de generación de variantes tanto en el documento como en el metatesauro. El sistema de anotación de discapacidades, como se puede apreciar en la Figura 1, se divide en varias fases.

Tenemos una fase inicial que consiste en la obtención del metatesauro con listas de discapacidades y con sus variantes para los términos involucrados. Este listado nos proporcionará la terminología básica para identificar las expresiones relativas a discapacidades en los textos. También se construye una colección de prueba de documentos del dominio médico. Partiendo de los nombres del conjunto de ERs descritas en la sección anterior, se ha hecho una búsqueda para recoger artículos científicos relacionados con ellas en los que es de esperar que se den apariciones de discapacidades.

Después para cada documento considerado se obtienen los sintagmas nominales (SN), los tokens de cada SN y las variantes de cada token. Se identifica también si los términos se corresponden con palabras de impedimento, que puedan ser un indicativo de discapacidad.

A continuación se establece la correspondencia entre los sintagmas nominales del documento (SND) y los candidatos del metatesauro. Esta relación será de 1 a N.

Después para las relaciones (SND-Candidatos) obtenidas en la fase anterior, se realiza un cálculo de afinidad, que permite establecer un ranking a partir del cual se selecciona el mejor candidato. La última fase es la evaluación del documento anotado por el sistema comparándolo con el correspondiente anotado manualmente.

A continuación, pasamos a analizar cada una de las fases del sistema una vez que la primera fase de adaptación de recursos se ha completado.

4.1 Tratamiento del documento y del metatesauro

Esta fase tiene como objetivo la obtención de una estructura que contenga la información de todos los SN obtenidos del documento y del metatesauro. Está compuesto de las siguientes fases:

- Obtención de cada una de las frases
- Obtención de los SN de cada frase
- Obtención de las variantes a partir de los SN

En esta fase utilizamos la herramienta de procesamiento de lenguaje natural OpenNLP [3]. Esta herramienta, permite detectar frases, realizar chunking, y análisis superficial. Nos quedamos con los sintagmas nominales más pequeños de los anidamientos proporcionados por el etiquetador.

El paso de la obtención de variantes requiere un procesamiento que consiste en:

- Filtro de palabras vacías.
- Tokenizador
- Filtro de puntuación
- Obtención de variantes de cada token.

Este proceso se realizará para cada SN que hayamos obtenido. Se inicia con un filtrado de palabras vacías. En este punto debemos mencionar que también se ha realizado un filtrado de nombres propios para evitar ruido en el sistema. Tras este primer filtro utilizamos el tokenizador proporcionado por la herramienta OpenNLP.

Uno de los puntos críticos del sistema es el de la obtención de variantes. En el sistema que estamos tratando, podemos seleccionar el nivel de obtención de variantes. Las variantes se generan de manera recursiva, por lo que este proceso, de obtención de variantes, va anidándose.

Las variantes que se generan podrán ser sinónimos o derivaciones. Tanto los sinónimos como las derivaciones son extraídos con la herramienta WordNet[4]. En la Figura 2 podremos ver un ejemplo gráfico de cómo se generan las variantes configurado a 2 niveles. En primer nivel, por cada uno de los tokens que

[3] https://opennlp.apache.org/
[4] https://wordnet.princeton.edu/

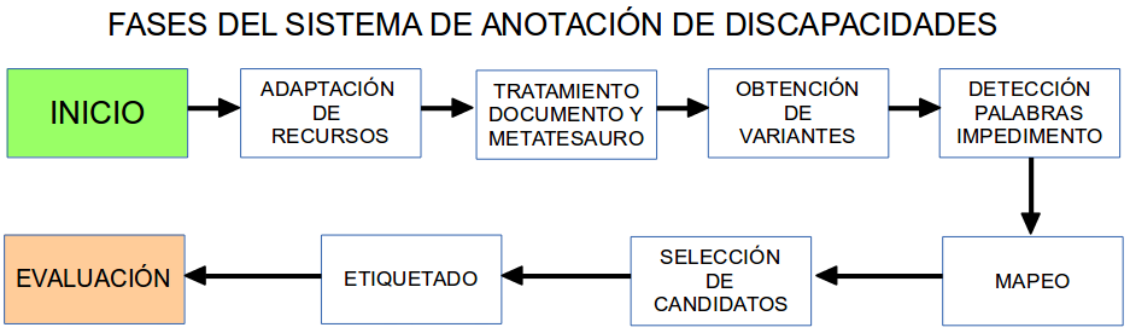

Figura 1: Arquitectura del sistema con las distintas fases de su ciclo de vida.

forman el SN se generan las diferentes variantes. En un segundo nivel, se vuelven a obtener las variantes a partir de las variantes obtenidas en el nivel anterior. La figura muestra el caso para el token *severe*.

La generación de variantes por niveles nos permite poder ampliar de manera dinámica la semántica de los SN que estamos tratando. Esto nos lleva al aumento del número de palabras y requiere un cálculo de la distancia semántica con respecto a la original. Con este fin, cada una de las variantes queda asociada a un historial: una cadena de caracteres que indica qué tipo de variante se ha generado en cada nivel recursivo para llegar a ésta. La posibilidad de poder configurar los niveles de generación de variantes tanto en el documento como en el metatesauro nos permite realizar un acercamiento semántico desde ambos extremos.

4.2 Palabras de impedimento

Se ha generado manualmente un conjunto de palabras, que parecen en la Tabla 1, que combinadas con funciones físicas o cognitivas da lugar a expresiones de discapacidad.

problems	lack	retardation
deterioration	impairment	failure
ataxia	worsening	disability
disablement	deficit	disorder
difficulty	deformity	loss
pain	abnormal	delayed
absent		

Tabla 1: Palabras de impedimento usadas en el sistema.

Las palabras de impedimento también son tratadas para generar variantes, pero en este caso a un único nivel. Posteriormente, en la fase de generación de candidatos, se comprueba la presencia de estas palabras en los SND que estemos tratando. De hecho no se considerará ningún SND que no contenga al menos una palabra relacionada con el impedimento.

4.3 Mapeo

Esta fase parte de dos fuentes de información: las variantes de cada SN del documento (SND) y las variantes del metatesauro de discapacidades.

El objetivo es la obtención de la relación entre SND-Candidatos. Dado un SND, se comprueba para cada una de sus palabras (no vacías) de cada una de sus variantes si se corresponde con alguna palabra de alguna variante de las expresiones de discapacidad del metatesauro (Candidatos). En el caso que exista al menos una correspondencia, la discapacidad que contiene la palabra estará entre los candidatos al SND. Con ello obtendremos un conjunto de estructuras de mapeos que recogen los candidatos asociados a cada SND (relación de 1 a N).

4.4 Selección de Candidatos

Este es el proceso principal del sistema. En esta fase se establecerán las puntuaciones de las relaciones entre los SND y los candidatos a evaluar. Las medidas principales que se establecen para realizar la evaluación son las siguientes: distancia, centralidad, cobertura y cohesión. El cálculo de estas medidas se ha inspirado en las utilizadas por Metamap (Aronson, 2001). Las medidas de distancia y centralidad se aplican a nivel de variantes. Mientras que, la cohesión y cobertura se aplican al nivel de SND y candidato. El proceso comienza con la evaluación de cada uno de los candidatos que se ha obtenido para cada una de las SND. Esto se calcula mediante la siguiente ecuación

$$generacion_{candidato} = \frac{distancia+centralizad}{v_c \times 6} + \frac{2\times(cobertura+cohesion)}{6} + impedimento$$

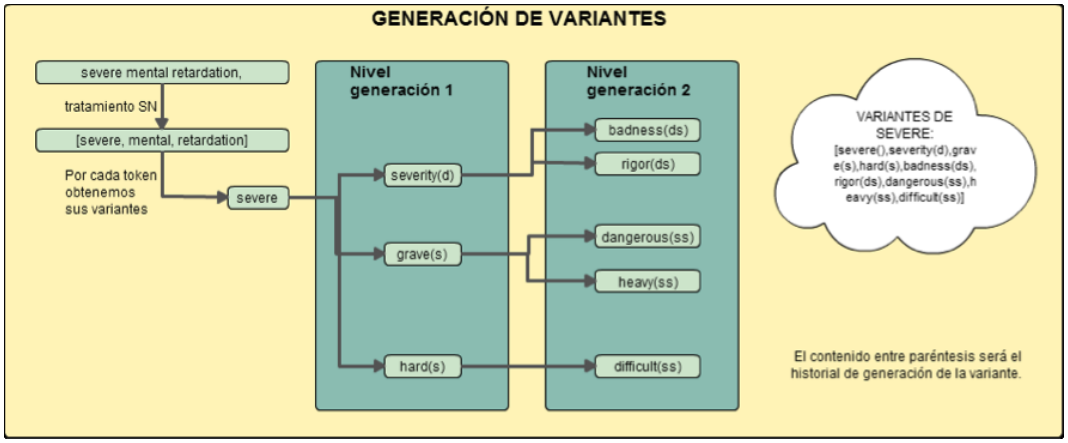

Figura 2: Generación de variantes.

donde v_e es el número de variantes del SND mapeado para cada token del candidato.

Por ejemplo, si el candidato tiene 3 tokens, puede haberse mapeado a 1,2 ó 3 variantes del SND como máximo. Para seleccionar la variante del SND que mejor se ajuste al token, se escoge aquella que maximice la relación entre las variantes

$selector_{variantes} = max(distancia_{xy} + centralidad_{xy})$

4.4.1 Cálculo de la distancia entre dos variantes

Consiste en obtener una valoración sobre la relación entre las variantes y los tokens originales. Para ello se realiza el cálculo de los pesos a partir del historial obtenido en la generación de las variantes.

$distancia_{xy} = \frac{base}{base + relacion_{xy}}$

donde

$relacion_{xy} = \frac{calculoHistorial(x) + calculoHistorial(y)}{2}$

y

$base = 4 + (nivelDocumento + nivelCandidatos) \times 3$

donde $nivelDocumento$ y $nivelCandidatos$ son los niveles de generación de variantes para el documento y para el metatesauro respectivamente. El cálculo del historial de una variante es la suma de los pesos establecidos dependiendo del tipo de generación de la variante. Para ello se recorre el historial de generación de la variante y se obtiene un valor que significará el coste de la generación de ésta como el indicado en la Tabla 2.

Tipo de Variante	Símbolo	Peso
Sinónimo	s	3
Derivación	d	1

Tabla 2: Peso establecido a cada tipo de variante.

Veamos un ejemplo del cálculo del historial en la Tabla 3.

Palabra	Historial	Peso
Severe		0
Severity	d	1
Badness	ds	4
dangerous	ss	6

Tabla 3: Ejemplo de calculo de un historial.

4.4.2 Cálculo de la centralidad de dos variantes

El cálculo de la centralidad entre dos variantes consiste en realizar la media entre la centralidad de la variante del metatesauro y de la variante del SND. La centralidad de una variante se basa en si el token original pertenece a la cabecera del árbol sintáctico. Si pertenece a la cabecera la centralidad para la variante será 1, en caso contrario 0.

$centralidad_{xy} = \frac{centralidad_x + centralidad_y}{2}$

4.4.3 Cálculo de la cobertura

Definimos cobertura como la medida que nos permite valorar el número de variantes coincidentes entre SND y candidato. Para ello lo obtendremos mediante la siguiente ecuación

$cobertura = \frac{cobertura_{SND} + 2 \times cobertura_{candidato}}{3}$

donde

$$cobertura_{SND} = \frac{v_e}{n_{tokens\_SND}}$$

$$cobertura_{candidato} = \frac{v_e}{n_{tokens\_candidato}}$$

donde v_e es el número de variantes mapeadas entre candidato y SND, $n_{tokensSND}$ el número de tokens que tiene el SND, $n_{tokenscandidato}$ el número de tokens que tiene el candidato. Tanto $cobertura_{SND}$ como $cobertura_{candidatos}$ pueden valer como máximo 1.

4.4.4 Cálculo de la cohesión

Definimos cohesión como la medida que nos permite valorar la conectividad entre las variantes del SND mapeadas con respecto a las conexiones posibles. Por ejemplo, si nos encontramos ante un caso en el que el número de variantes mapeadas como máximo fuese 7, podrían existir 6 conexiones. Esta medida calcula la relación entre el número de conexiones que contiene el mapeo con respecto al máximo de conexiones posibles que podría tener.

$$coherencia = \frac{conexiones^2_{existentes}}{conexiones^2_{maximas}}$$

En el caso que tengamos un mapeo en el que no se pueda realizar conexiones se establecerá este valor a 0.

4.5 Etiquetado

Una vez evaluadas todas las relaciones SND-Candidato, se procede a realizar el proceso de etiquetado del documento. Para realizar este proceso, se recorre cada una de las frases del documento y se comprueba si los SND de la frase contiene algún candidato asociado. En caso afirmativo, se comprueba que esa relación SND-Candidato tenga un valor mayor o igual a un umbral que establecemos. El umbral que tenemos establecido actualmente en el sistema es de '0.6'. Por tanto, toda relación SND-Candidato que supere o iguale ese umbral, quedará reflejada en el documento de salida.

5 Resultados

Hemos calculado los resultados de precisión, cobertura y medida-F obtenidos con diferentes configuraciones. La Tabla 4 muestra los resultados para distintos niveles de cálculo de variantes cuando no se utilizan las palabras de impedimento. Hemos generado un baseline en el que no se utilizan variantes y en el que se detecta una discapacidad tan solo si se encuentran términos en el sintagma nominal analizado que se correspondan con alguna entrada del metatesauro. La Tabla 5 muestra los resultados para distintos niveles de cálculo de variantes utilizando las palabras de impedimento. Comparando ambas tablas observamos que la introducción de las palabras de impedimento supone una mejora notable de los resultados. Lógicamente aumenta la cobertura, pero también aumenta la precisión. Esto se debe a que la introducción del peso correspondiente a la presencia de las palabras de impedimento en la valoración de un candidato cambia el ranking y produce distintas asociaciones.

Nivel	Medida-F	P	C
Baseline	0.41	0.68	0.33
Nivel 1-1	0.42	0.57	0.37
Nivel 2-1	0.30	0.25	0.42
Nivel 3-1	0.19	0.12	0.48
Nivel 2-2	0.23	0.16	0.47

Tabla 4: Resultados en función de la Medida-F, Precisión (P) y Cobertura (C) para diferentes niveles de variación sin utilizar palabras de impedimento.

Centrándonos ya la Tabla 5, vemos que los mejores resultados se obtienen para el caso (3-1) y (2-2), al considerar dos y tres niveles de generación de variantes desde los sintagmas nominales y un nivel o dos desde el tesauro. La pérdida de precisión que se produce para más niveles no compensa la ganancia en cobertura.

Nivel	Medida-F	P	C
Baseline	0.40	0.82	0.29
Nivel 1-1	0.45	0.88	0.33
Nivel 2-1	0.49	0.86	0.36
Nivel 3-1	0.51	0.76	0.41
Nivel 2-2	0.51	0.78	0.41

Tabla 5: Resultados en función de la Medida-F, Precisión (P) y Cobertura (C) para diferentes niveles de variación.

Aunque estos resultados son preliminares, consideramos que son un buen comienzo para abordar la tarea. Así, algunos estudios (Gonzalez y Iglesias, 2011) indican resultados de precisión de Metamap de entre 40 y 45 % y de medida-F de 20 % para conceptos médicos recogidos en UMLS. De esta forma y a pe-

sar de que los resultados no son comparables, muestra una idea general de cómo son los resultados obtenidos.

6 Conclusiones y Trabajos Futuros

Nuestra propuesta para anotar discapacidades en documentos médicos se basa en la generación de variantes de la lista de discapacidades consideradas. Para ello, se ha elaborado un metatesauro específico de discapacidades sobre el cual se ha aplicado la generación de variantes. Hasta el momento las variantes consideradas han sido derivacionales y sinónimos. Con respecto a MetaMap, que es la herramienta de referencia en este tipo de tareas, hemos incluido varias mejoras tratando de adaptar la identificación de conceptos médicos al problema específico de las discapacidades. No obstante, estas mejoras pueden ser aplicadas a otros tipos de conceptos médicos. El sistema propuesto, permite también configurar el nivel de generación de variantes, tanto en los documentos analizados como en el metatesauro. Otra mejora, ha sido el uso de palabras de impedimento, que nos ha permitido expandir el conjunto de discapacidades consideradas, y convertir el listado de funciones corporales y cognitivas de Orphanet en una metatesauro de discapacidades. Los primeros resultados obtenidos indican que el sistema es capaz de obtener niveles competitivos de precisión y cobertura.

Sin embargo, se presentan muchas posibilidades de mejorar los resultados. En primer lugar queremos ampliar el corpus de prueba utilizado. También ampliaremos la lista de palabras de impedimento, analizando el efecto que pueda tener la inclusión de cada una de ellas. Así mismo nos proponemos ampliar las técnicas de PLN utilizadas, incluyendo tratamiento de la negación y desambiguación.

Bibliografía

Aronson, Alan R. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. En *Proceedings of AMIA, Annual Symposium*, páginas 17–21.

Carrero, F. M., J. C. Cortizo, J. M. Gómez, y M. de Buenaga. 2008. In the development of a spanish metamap. En *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, páginas 1465–1466, New York, NY, USA. ACM.

Castro, E., A. Iglesias, P. Martínez, y L. Castaño. 2010. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. En *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, páginas 751–757, New York, NY, USA. ACM.

Conrado, M. S, W. Koza, J. Dıaz-Labrador, J. Abaitua, Solange O Rezende, T. AS Pardo, y Z. Solana. 2011. Experiments on term extraction using noun phrase subclassifications. páginas 746–751.

Gonzalez, R.Paula F. y E. L. Iglesias. 2011. Study and evaluation of an indexing tool: Metamap. Informe Técnico TFM_SSIA_2010-11, Universidad de Vigo.

Iglesias, A., E. Castro, R. Perez-Lainez, L. Castaño, P. Martínez, J. M. Gómez-Pérez, S. Kohler, y R. Melero. 2008. MOSTAS: un etiquetador morfo-semántico, anonimizador y corrector de historiales clínicos. *Procesamiento del Lenguaje Natural*, 41.

Organization., World Health. 2007. *nternational classification of functioning, disability and health : children and youth.* World Health Organization Geneva.

Oronoz, M., A. Casillas, K. Gojenola, y A. Perez. 2013. Automatic annotation of medical records in spanish with disease, drug and substance names. En *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volumen 8259 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, páginas 536–543.

Oronoz, M., A. Díaz de Ilarraza, y O. Torices. 2010. First steps in the manual and automatic annotation of clinical notes in spanish. *Procesamiento del Lenguaje Natural*, 45:259–262.

Vivaldi, J. y H. Rodríguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45:251–254.