Desarrollo de un diccionario electrónico para la detección automática de candidatos a término del dominio médico. Una aplicación con Smorph y MPS

Walter Koza, María José Mánquez Contreras, Mirian Muñoz Araya
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso, Av. El Bosque 1290, CP 2530388. Proyecto Fondecyt 11130469
walter.koza@ucv.cl
mariajose.manquez1@gmail.com
m.aracely.m.a@gmail.com

Resumen. El presente trabajo, enmarcado en el proyecto FONDECYT 11130469, describe la metodología de la elaboración de un diccionario médico electrónico, con el propósito de ser implementado en tareas de extracción automática. Se compilaron los lemas incluidos en Diccionario Mosby [1] y Terminología Médica [2] que se correspondían con las estructuras (i) nombre; (ii) nombre – adjetivo, y (iii) nombre – preposición de – nombre. Cada nombre y adjetivo fue cargado de manera separada y se le asignó a cada uno un modelo correspondiente con sus rasgos morfológicos. Posteriormente, se establecieron reglas de reagrupamiento a fin de conformar las estructuras (ii) y (iii). Para la implantación en máquina, se recurrió a los software Smorph [3] y Módulo Post Smorph [4]. Smorph permite analizar morfológicamente la cadena de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada ocurrencia de acuerdo con los rasgos declarados. MPS, por su parte, toma como input el output de Smorph y, a partir de reglas de recomposición, descomposición y correspondencia declaradas por el usuario, analiza la cadena de lemas resultante del análisis morfológico.

Palabras clave: diccionario médico, término médico, extracción automática, Smorph, Módulo Post Smorph

1 Introducción

El presente trabajo describe la metodología desarrollada en el proyecto Fondecyt 11130469 para la obtención de un extractor de candidatos a término del dominio médico mediante el procesamiento de información lingüística. En este caso, se trabajó con reglas en el nivel léxico, morfológico y sintáctico. Esto se enmarca en el campo de la lingüística computacional, en general, y, en particular, en tareas de minería textual. Tales efectos, se toman en consideración dos tipos de antecedentes: las investigaciones sobre terminología y extracción terminológica, y el uso de formalismos y software declarativo. El presente trabajo complementa investigaciones previas [5], [6], [7].

Dado el creciente desarrollo de las tecnologías de la comunicación, que involucra una gran producción y circulación del conocimiento científico es necesario contar con sistemas capaces de procesar grandes cantidades de datos con los que se enfrentan los usuarios diariamente. A tales efectos, para acceder a esa gran masa de datos, se hace necesario disponer de herramientas que puedan procesarlos y que cuenten con sistemas de almacenamiento y de recuperación de la información [8]. Al mismo tiempo, también resulta fundamental desarrollar recursos que regulen y analicen los conceptos de las distintas áreas del conocimiento, así como también la asignación de denominaciones nuevas para los nuevos conceptos que están surgiendo, con el objetivo de garantizar una adecuada comunicabilidad científica. Una de las actividades principales en el desarrollo de dichos sistemas es la detección automática de términos de dominios específicos. En este sentido, un término es una unidad léxica que designa a un concepto en un campo temático particular [9], [10], a la vez, desde la perspectiva de la lingüística de corpus, se puede considerar término al output de un proceso terminológico [11].

La extracción de términos representativos de un área suele constituir el punto de partida para realizar tareas más complejas, como ser la elaboración de listas de entradas para diccionarios especializados, creación de base de datos o de ontologías y taxonomías, etcétera. Entre los inconvenientes principales, se encuentra el cambio constante de la terminología, lo que impide mantener bases terminológicas actualizadas inmediatamente por medios manuales e implica la necesidad de herramientas que puedan detectar tanto los

términos nuevos que se creen, así como también las variaciones que puedan observarse en ellos [12]. Por otro lado, las tareas de extracción, sobre todo las que apelan a técnicas de análisis lingüístico, suelen enfocarse en áreas de conocimiento específicas, con el objeto de adaptarse a los requerimientos y particularidades propias de cada una de ellas.

Ahora bien, una de las áreas fundamentales es la de la medicina, no solo por la función social que cumple, conservar la integridad física de los seres humanos, sino también por la creciente producción y circulación de textos del área (artículos, casos clínicos, informes, etcétera). A tales efectos, el presente trabajo propone el desarrollo de una herramienta que auxilie en este tipo de tareas.

De acuerdo con Cabré [13], la complejidad que entraña la detección automática de términos implicaría el desarrollo de un procesador con las mismas habilidades de un especialista humano; dicha postura podría resultar extrema en la medida en que sería imposible dotar a un extractor con dichas habilidades. No obstante, es posible que las máquinas procesen algo de la misma información que los especialistas; se trataría de información léxica, morfológica y sintáctica. A tales efectos, se propone la siguiente definición de término médico:

> Dado un corpus compuesto por textos del dominio médico, un término médico es un sintagma, generalmente nominal, que posee un significado que puede ser adjudicado a dicha área y al que es posible acceder a través de un proceso de detección automática basado en información lingüística porque:
>
> i. Su lema es una entrada léxica de un diccionario electrónico del dominio médico.
> ii. Posee una estructura morfológica propia del dominio médico que se puede formalizar y ser implantada en máquina.
> iii. Incluye un neologismo que no posee una estructura morfológica, pero su categoría gramatical puede ser deducida automáticamente a través del contexto sintáctico [7].

SMORPH es un utilitario que agrupa un conjunto de funcionalidades en torno a la morfología: compilación de diccionario (transformación de un diccionario fuente, es decir, una descripción morfológica legible de un conjunto de palabras, en una representación interna utilizable para el análisis y/o generación morfológica), análisis y generación morfológicas, segmentación y lematización de textos. Puede ser considerado como un elemento de un sistema de tratamiento lingüístico más amplio en el que su papel principal sería asegurar la primera parte del tratamiento: convertir los archivos de textos ASCII a secuencias de ítems significativos adecuados como entrada del análisis sintáctico. En este cuadro, SMORPH permite construir diccionarios electrónicos voluminosos necesarios para un análisis lingüístico de textos. Estos diccionarios podrán utilizarse en generación o en segmentación y análisis. [3]

A tales efectos, en el marco del proyecto Fondecyt nº 11130469, se está desarrollando un extractor de candidatos a término del dominio médico para el español a partir de técnicas lingüísticas. Para ello se elaboran reglas que contienen información léxica, morfológica y sintáctica [7]. En relación con el primer nivel, se elaboró un diccionario electrónico que contuviera términos médicos. Para la elaboración de dicho diccionario, se recurrió a los lemas incluidos en la versión electrónica del diccionario *Mosby* [1] y *Terminología Médica* [2] que conformaban un total de 27.923 lemas correspondientes con nombres (comunes y propios) y adjetivos. Posteriormente se estableció una descripción de los rasgos morfológicos de estas expresiones y se crearon modelos a partir de las regularidades detectadas. Dicha descripción fue implantada en máquina. El trabajo computacional se realizó con las herramientas Smorph [3] y Módulo Post Smorph [4]. El primero permite analizar morfológicamente la cadena de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada ocurrencia de acuerdo con los rasgos declarados. MPS, por su parte, tiene como input la salida de Smorph y, a partir de reglas de recomposición, descomposición y correspondencia declaradas por el usuario, analiza la cadena de lemas resultante del análisis morfológico.

El trabajo se organiza de la siguiente manera. En la sección 2, se presentan los antecedentes de las tareas de extracción automática en el dominio médico. En la sección 3, se describe la metodología y el trabajo realizado. En la sección 4, se discuten los resultados obtenidos. Finalmente, en la sección 5, se presentan las proyecciones y limitaciones de la propuesta.

2 Extracción de términos en el área médica

En el campo médico, Krauthamer y Nenadić [12] mencionan que las barreras para una extracción de términos exitosa incluyen problemas como las variaciones léxicas, la sinonimia y la homonimia. Por otro lado, el mantenimiento de los recursos terminológicos se dificulta ante el constante cambio de la terminología, algunos términos aparecen solo por un período corto de tiempo y se introducen nuevos en el vocabulario del dominio, prácticamente, a diario. A la vez, a eso hay que sumarle la falta de convenciones firmes en la nomenclatura, pues, si bien existen directrices para algunos tipos de entidades médicas, estas no imponen restricciones a los expertos del dominio, quienes no están de ningún modo obligados a usarlas cuando se acuña un nuevo término. Consecuentemente, junto con los términos "bien formados" existen nombres ad-hoc, los cuales son problemáticos para los sistemas de identificación de términos. No obstante, a pesar de las dificultades mencionadas, se han venido desarrollando diversos sistemas de reconocimiento de términos para muchas clases de entidades médicas. Estos se basan tanto en características internas de clases específicas o en pistas externas que pueden ayudar al reconocimiento de secuencias de palabras que representan conceptos del dominio. Para ello, se utilizan diferentes tipos de características, tales como ortografía (mayúsculas, dígitos, caracteres griegos) y pistas morfológicas (afijos específicos y formantes cultos) o información proveniente del análisis sintáctico. Además, se sugieren diferentes medidas estadísticas para promover candidatos a términos a términos.

Para el caso del español, pueden mencionarse los trabajos realizados por López, Tercedor y Faber [14], para el proyecto Oncoterm. Se trata de una investigación interdisciplinar sobre terminología con el propósito de elaborar un sistema de información sobre el subdominio médico de la oncología en donde los conceptos se vinculen a una ontología. Para ello, recurren a información extraída de diccionarios y de corpus textuales especializados como así también proporcionada por expertos.

Castro y sus colaboradores [15], por su parte, presentan una propuesta para la detección de conceptos de notas clínicas, implementando una herramienta para la identificación de conceptos biomédicos en la ontología SNOMED CT. Para ello, describen el proceso de anotación semántica de los términos de dicha ontología en un corpus compuesto por notas clínicas. Los experimentos se centraron en ver qué tan estrechamente el etiquetado automático de conceptos que realiza SNOMED CT se refleja en la anotación manual llevada a cabo por expertos del área. De acuerdo con los autores, las funcionalidades de la herramienta permiten la obtención de un mayor conocimiento semántico, que influyen en el establecimiento de nuevas relaciones que permitan la minería de texto en las notas clínicas.

A su vez, tomando como base SNOMED CT y otras ontologías como UMLS, se han realizado estudios de reconocimiento automático de similitud semántica. Entre ellos, se pueden mencionar los llevados a cabo por Sánchez, Batet y Valls [18], y Garla y Brandt [19]. Ambos trabajos están enfocados en analizar automáticamente la relación entre conceptos que comparten el mismo contexto.

Por otro lado, recurriendo a información semántica extraída de la Wikipedia, Vivaldi y Rodríguez [20] presentan un sistema de extracción de términos probado en un corpus médico. Los experimentos consisten en tomar un documento y el correspondiente conjunto de candidatos a términos y comparar los resultados que se obtienen recurriendo a EuroWordNet y Wikipedia. Esto consiste en explorar el segundo recurso con el fin de obtener un coeficiente de dominio equivalente al obtenido con EuroWordNet. Este método, consiste en, para un candidato a término dado, (i) encontrar una página de Wikipedia que se corresponda con este, (ii) encontrar todas las categorías de Wikipedia asociadas a tal página, y, por último, (iii) explorar la Wikipedia siguiendo recursivamente todos los links de categorías encontrados en (ii) a fin de enriquecer el borde de dominio. Según los autores, los resultados demuestran que este recurso puede utilizarse para tareas de extracción automática de términos.

Por último, ya en el ámbito de la traducción y la lingüística de corpus, Moreno-Sandoval y Campillo-Llanos [21] elaboran un corpus compuesto por textos biomédicos en español, árabe y japonés. Los textos incluidos en dicho corpus no son extremadamente técnicos, sino dirigidos a estudiantes de medicina, como, por ejemplo, manuales, y revistas médicas destinadas al público en general. El propósito de los autores es desarrollar un buscador de términos en dicho corpus para las tres lenguas y poder compararlas.

En lo que atañe a los métodos basados exclusivamente en el procesamiento de información lingüística, estos pueden dividirse en dos enfoques: los basados en diccionarios y los basados en reglas morfológicas y sintácticas.

Por un lado, los métodos constituidos a partir de diccionarios utilizan recursos terminológicos existentes con el propósito de localizar las ocurrencias de términos en los textos. La limitación, obvia, que presentan es que muchas ocurrencias pueden no ser reconocidas si se recurre a diccionarios o bases de datos estándares, no obstante, en el presente trabajo, se puede apreciar que contar con la información lexicográfica de los diccionarios proporciona una base idónea para las tareas de extracción de términos. Por otro lado, también puede influir negativamente factores como la homonimia y las variaciones en el deletreado de los términos, por ejemplo, variaciones en la puntuación (*bmp-4*/*bmp4*), uso de diferentes numerales (*syt4*/*syt iv*), diferencias en la transcripción de letras del alfabeto griego (*igα*/*ig alpha*) o variaciones en el orden (*integrin alpha 4*/*integrin4 alpha*) [22].

Por otro lado, los enfoques basados en reglas morfológicas, por su parte, intentan recuperar términos por el restablecimiento asociado a los patrones de formación que han sido utilizados para construir los términos en cuestión. Se trata de desarrollar reglas que describan las estructuras de denominación común para ciertas clases de términos usando pistas ortográficas o léxicas, como así también, características morfosintácticas más complejas. Desde esta perspectiva, se puede mencionar el trabajo de Segura, Martínez y Sami [23], focalizado en la detección automática de fármacos genéricos mediante la utilización del metatesauro ULMS y reglas de nomenclatura para la formación de fármacos genéricos propuestas por el consejo United States Adoptated Names (USAN), el cual permite la clasificación de los fármacos en familias farmacológicas. Con esta técnica, se pueden detectar fármacos no incluidos en UMLS. Los autores logran un 100% de cobertura y un 97% de precisión utilizando UMLS, y 99,3% de precisión y un 99,8% de cobertura recurriendo a una combinación de información lexicográfica propuesta por UMLS y reglas de formación de nombres de fármacos propuestas por USAN. Posteriormente, Gálvez [25] propone un trabajo similar aunque basado solamente en reglas morfológicas, al igual que Segura, Martínez y Sami [23], propuestas por USAN, y recurriendo a la herramienta de estados finitos NooJ. De esta manera, la autora logra 99,8% de precisión y 92% de cobertura.

Como se mencionó más arriba, para el desarrollo del extractor automático de candidatos a término del dominio médico se pretenden establecer reglas en el nivel léxico, morfológico y sintáctico. En el presente artículo, se describen las tareas realizadas en el primer nivel.

3 Metodología propuesta

3.1 Software utilizado

Para el trabajo de extracción automática mediante el diccionario médico implantado, se recurrió a los softwares Smorph [3] y Módulo Post Smorph (MPS) [4], que trabajan en bloque. El primero es un analizador y generador textual que en una sola etapa realizar la segmentación, lematización y análisis morfológico, léxico, semántico, etcétera, según los rasgos declarados por el usuario. MPS, por su parte, toma como input el output de Smorph y mediante reglas de reagrupamiento, descomposición y correspondencia, también declaradas por el usuario, analiza la cadena del análisis resultante de Smorph. A continuación se describen brevemente ambos programas.

3.1.1 Smorph

SMORPH (Segmentación y Morfología) reduce las tres etapas de normalización, segmentación y lematización a una sola. Es un analizador y generador textual que en un único paso realiza la delimitación previa de los segmentos textuales a considerar (tokenización) y el análisis morfológico (lematización) dando como resultado las formas pertenecientes a un lema con los valores correspondientes. Se trata de una herramienta declarativa, lo que implica que la información utilizada está separada de la maquinaria algorítmica. Esto hace que se la pueda adaptar al uso que quiera darse, ya que con el mismo software se puede tratar cualquier lengua si le cambia la información lingüística. Su autor lo describe de la siguiente manera:

> SMORPH es un utilitario que agrupa un conjunto de funcionalidades en torno a la morfología: compilación de diccionario (transformación de un diccionario fuente, es decir, una descripción morfológica legible de un conjunto de palabras, en una representación interna utilizable para el análisis y/o generación morfológica), análisis y generación morfológicas, segmentación y lematización de textos. Puede ser considerado como un elemento de un sistema de tratamiento lingüístico más amplio en el que su papel principal sería asegurar la primera parte del tratamiento: convertir los archivos de textos ASCII a secuencias de ítems significativos adecuados como entrada del análisis sintáctico. En este cuadro, SMORPH permite construir diccionarios electrónicos voluminosos necesarios para un análisis lingüístico de textos. Estos diccionarios podrán utilizarse en generación o en segmentación y análisis. [3]

En este programa se declaran cinco tipos de información: códigos ascii, rasgos, terminaciones, modelos y entradas. Para una mejor comprensión, se invertirá el orden de la explicación, comenzando por indicar lo que se declara en:

Entradas: Dentro de este sistema, las entradas son, en realidad, el diccionario lingüístico, un diccionario especial en el que las expresiones (palabras) tienen la posibilidad de aparecer:

A partir de los lemas con la indicación precisa del modelo morfológico que siguen **(1)**

```
Médico  @nmed3 .
Cofre   @n2 .
```

Directamente con la indicación de los rasgos morfológicos **(2)**

```
De      /prep .
```

En el caso de ‘médico’ se presenta el lema que se expresa convencionalmente con la forma masculina singular, como ocurre en los diccionarios comunes. Es decir, ‘médico’ es el lema que representa al grupo de sustantivos médicos ‘médico’, ‘médica’, ‘médicos’, ‘médicas’. ‘Cofre’, por su parte, es un sustantivo que pertenece al grupo de sustantivos 2. En el caso de la preposición ‘de’, no se recurre a ningún modelo, sino que solo se señala el carácter de preposición mediante la expresión ‘prep’. Para este trabajo, se armaron modelos para sustantivos y adjetivos médicos y en el caso de las siglas que están registradas en los diccionarios, se cargaron solo con la expresión ‘abmed’.

Modelos: En los modelos, se consigna la estructura morfológica. Los modelos vienen introducidos por @, que indica el lugar en que va la forma básica o raíz a la que se concatenan las terminaciones.

```
@nmed           -0
        +@      nomed/fem/sg
        +s      nomed/fem/pl .

@nmed2          -0
        +@      nmed/masc/sg
        +s      nmed/masc/pl .

@nmed3          -1
        +o      nmed/masc/sg
        +os     nmed/masc/pl .
        +a      nomed/fem/sg
        +as     nomed/fem/pl .
```

Primero se indica el número de caracteres que se extrae al lema. Así, en el primer ejemplo, no se extrae ningún carácter, podría tratarse de ‘cama’ ya que para formar el singular se deja sin modificaciones (@ en la

terminación significa terminación vacía) y para formar plural se agrega ‘s’. Las dos terminaciones son acompañadas por la lista de rasgos/valores.

El segundo ejemplo de modelo podría asignarse a ‘diagnóstico’, tiene la misma formación que ‘cama’, pero, en lugar de tener un rasgo femenino, lo tiene masculino. El tercer ejemplo, correspondiente a ‘médico’, entre otros, resta un carácter del lema (queda ‘medic’) y agrega las distintas terminaciones (o, os, a, as). Los modelos especifican ocurrencias complejas pro medio de la concatenación de cadenas adyacentes.

Terminaciones: Se trata de serie de caracteres que expresan un rasgo o un conjunto de rasgos. Ejemplos: o, a, as, os, ción, ciones, e, es, etcétera.

Rasgos: Para construir los modelos, se recurre a rasgos morfológico-sintácticos y, en esta ocasión, a la información léxica médica. De este modo, por ejemplo, se tienen:

EMS (etiqueta morfosintáctica), que incluye los valores ‘n’ (nombre), ‘nmed’ (nombre médico), ‘adj’ (adjetivo), ‘adjmed’ (adjetivo médico), ‘v’ (verbo), ‘adv’ (adverbio).
GEN (género), que incluye ‘masc’ (masculino), ‘fem’ (femenino), los casos neutros están indicados con __
NUM, que incluye ‘sg’ (singular), ‘pl’ (plural’.

ASCII: En la descripción que sobre Smorph proporciona Aït Mokhtar [3] señala que el diccionario se realiza a través de un autómata de estados finitos y sus archivos fuentes usan editores de textos planos que describen definiciones tipográficas sobre los caracteres, terminaciones, modelos de flexión y entradas léxicas. Las definiciones ASCII sobre la información tipográfica indican las clases de separadores, espacios y potenciales tipográficos. Estos últimos indican el conjunto de representaciones que puede asumir un carácter en ese tipo de texto, por ejemplo de A podemos declarar a, à, á, Á, À como representaciones posibles. Las terminaciones serán utilizadas en los modelos de flexión y los rasgos, en las definiciones morfosintácticas asociadas a las formas. A la vez, también se puede declarar para cada rasgo el conjunto de sus valores posibles. Cada modelo de flexión describe las flexiones que pueden darse de las distintas categorías, como así también los rasgos morfológicos correspondientes.

3.1.2. Módulo Post Smorph (MPS)

MPS realiza tratamientos previos a los de la sintaxis general de la oración, con el objetivo de normalizar la entrada de la sintaxis estándar, como ser fechas, cantidades, cuestiones relativas a la sufijación y prefijación, el tratamiento de los clíticos y de las contracciones. Este programa, al igual que Smorph, también es una herramienta declarativa, con la que, mediante ciertas reglas, se pueden expresar los valores de entradas (sobre dos o más estructuras de datos de la salida de Smorph) y los valores de salida sobre la estructura reagrupada. En las indicaciones sobre el trabajo de MPS, Bès y Solana [27] señalan que MPS compara una ocurrencia analizada (oa) en la entrada a analizar con el primer elemento de c/una de sus reglas. Si encuentra una regla cuyo primer elemento subsume el oa, verifica el oa siguiente con el elemento siguiente en la entrada de la regla, y así siguiendo. MPS aplica la primera regla en cuya entrada todos los elementos subsumen un oa en la entrada a analizar.

Mientras que en Smorph es preciso declarar cinco tipos de información, en MPS las fuentes declarativas se constituyen con un único archivo: rcm. Rcm es un listado de reglas de reagrupamiento, descomposición y correspondencia que especifica cadenas posibles de lemas a partir de una sintaxis determinada para el software. Las reglas pueden ser de tres tipos:

De reagrupamiento. Se trata de juntar varios lemas (palabras, signos de puntuación, etcétera) y agruparlos bajo una misma etiqueta. Artículo + Nombre = SN **(3)**
De descomposición. Sería lo inverso del reagrupamiento. Contracción = Preposición + Artículo **(4)**
De correspondencia. Artículo = Determinante **(5)**

En la sección siguiente, se detalla la metodología realizada para la extracción.

3.2 Detección automática de candidatos

En esta ocasión, se elaboraron reglas de reagrupamiento y de correspondencia para la identificación automática de términos. Se trabajó con candidatos a término que se correspondieran con las siguientes estructuras:

Unigramas: Nombre (‘cáncer’) **(6)**
Bigramas: Nombre + Adjetivo (‘cáncer mamario’) **(7)**
Trigramas: Nombre + preposición ‘de’ + Nombre (‘cáncer de mama’) **(8)**

Para ello, se combinaron los siguientes elementos: nombre (N), nombre médico (NM), nombre propio médico (NPM), adjetivo (A), adjetivo médico (AM), preposición ‘de’ y lo que se denominó palabras desconocidas (‘PD’), es decir, aquellas expresiones consideradas neologismos por no estar en el diccionario del software analizador. En el presente trabajo se asume que las PD que pueden categorizarse automáticamente (mediante estructura morfológica o contexto sintáctico) son, en su mayoría, términos o partes de términos del dominio médico. En la tabla se muestran las reglas de reagrupamiento establecidas para la detección automática de los candidatos.

Tabla1. Estructura de las expresiones extraídas para la elaboración de listas de referencia

Unigramas	Bigramas	Trigramas
NM	NM + AM	NM + de + NM
NPM	NM + A	NM + de + N
PD	NM + NM	NM + de + PD
	NM + PD	N + de + NM
	N + AM	N + de + N
	N + PD	N + de + PD
	PD + AM	PD + de + NM
	PD + A	PD + de + N
	PD + N	PD + de + PD
	NM + NM	
	NM + N	
	N + N	
	NM + PD	
	NE + PD	

Una vez realizada la extracción inicial, se establecieron unas primeras reglas, también de reagrupamiento, con el propósito de aceptar o desechar los candidatos que involucraban PD. Dicho método se probó en dos corpus de textos médicos compuestos por dos géneros, casos clínicos médicos y artículos de investigación científica médica.

En el primer caso, se trabajó con el corpus CCM-2009, compilado por Burdiles [28], el cual está conformado por casos clínicos médicos aparecidos entre los años 1999 y 2008 en revistas médicas chilenas indizadas. Dicho corpus reúne un total de 99 textos y 80.224 palabras. Los casos clínicos reunidos pertenecen a las especialidades de Parasitología, Neuropsiquiatría, Enfermedades Respiratorias, Otorrinonaringología y Cirugía, Infectología, Pediatría, Obstetricia, Cirugía y Ciencias Biomédicas y Medicina Interna y Especialidades derivadas.

El segundo, denominado AICM-2014, fue compilado durante el 2014. Contiene artículos de investigación aparecidos entre los años 2001 y 2013, suma un total de 305 textos y 1.333.28 de palabras. Incluye las especialidades de Cardiología, Farmacología, Neurología y Oncología.

La extracción automática de candidatos a término se realizó de la siguiente manera:

Etapa I: Análisis morfológico a partir de Smorph y asignación de la etiqueta ‘med’ a los nombres y adjetivos del corpus que se encontraban en el listado de lemas. En esta etapa se le asignó la etiqueta PD a los neologismos.

Etapa II: Extracción de unigramas, bigramas y trigramas que involucrasen ‘med’ o ‘PD’.
Etapa III: Evaluación de resultados.

A modo de ejemplo, se presenta la extracción realizada en un fragmento del corpus. En la figura, se muestra el texto extraído de un artículo de investigación científica médica.

> Es importante enfatizar que los efectos de las estatinas no se limitarían a patologías que afectan el sistema cardiovascular.

Fig. 1. Fragmento del corpus AICM

A continuación se presenta el etiquetado de la primera oración mediante Smorph:

> 'Es'. ['ser', 'EMS','v', 'EMS','ind', 'PERS','3a', 'NUM','sg', 'TPO','pres', 'TDIAL','estrpi']. 'importante'. ['importante', 'EMS','adj', 'GEN','_', 'NUM','sg']. 'enfatizar'. ['enfatizar', 'EMS','v', 'EMS','infin', 'TR','r', 'TC','c']. 'que'. ['que', 'EMS','rel']. ['que', 'EMS','sub']. 'los'. ['el', 'EMS','det', 'TDET','art']. ['lo', 'EMS','cl', 'TPCL','nrfl']. 'efectos'. ['efecto', 'EMS','nom', 'GEN','masc', 'NUM','pl']. **['efecto', 'EMS','nommed', 'GEN','masc', 'NUM','pl'].** 'de'. ['de', 'EMS','prep']. ['de', 'EMS','prde']. 'las'. ['el', 'EMS','det', 'TDET','art']. ['lo', 'EMS','cl', 'TPCL','nrfl']. **'estatinas'. ['estatinas', 'EMS', 'PD'].** 'no'. ['no', 'EMS','advneg']. 'se'. ['lo', 'EMS','cl', 'TPCRF','rflse']. 'limitarían'. ['limitar', 'EMS','v', 'EMS','ind', 'PERS','3a', 'NUM','pl', 'TPO','cond', 'TR','r', 'TC','c', 'TDIAL','estrpi']. 'a'. ['a', 'EMS','prep']. 'patologías'. **['patología', 'EMS','nommed', 'GEN','fem', 'NUM','pl'].** 'que'. ['que', 'EMS','rel']. ['que', 'EMS','sub']. 'afectan'. ['afectar', 'EMS','v', 'EMS','ind', 'PERS','3a', 'NUM','pl', 'TPO','pres', 'TR','r', 'TC','c', 'TDIAL','estrpi']. 'el'. ['el', 'EMS','det', 'TDET','art']. 'sistema'. ['sistema', 'EMS','nom', 'GEN','masc', 'NUM','sg']. ['sistema', 'EMS','nom', 'GEN','masc', 'NUM','sg']. **['sistema', 'EMS','nommed', 'GEN','fem', 'NUM','sg'].** 'cardiovascular'. **['cardiovascular', 'EMS','adjmed', 'GEN','_', 'NUM','sg'].** '.'. ['pf', 'EMS','pun'].

Fig. 2. Output de Smorph

Como se puede apreciar, el software detecta ‘efectos’, ‘patologías’, ‘sistema’ y ‘cardiovascular’ como nombres y adjetivos del dominio médico. Asimismo, cabe observar que ‘efectos’ y ‘sistema’ poseen un doble etiquetado puesto que también son palabras de uso común. Por otro lado, ‘estatinas’ fue etiquetada como PD puesto que no se encuentra en el diccionario de Smorph, aunque dado lo que se asume en el presente trabajo, se establecería como candidato.

Una vez realizado el análisis morfológico, se tomó la salida de Smorph como input y se obtuvieron los siguientes candidatos a término:

Unigramas:
['efecto', 'EMS', 'CT_Unimed_DN'].
['patología', 'EMS', 'CT_Unimed_DN'].
['estatinas', 'EMS', 'CT_Unimed_PD'].
['reducción', 'EMS', 'CT_Unimed_DN'].
['evolución', 'EMS', 'CT_Unimed_DN'].
['serie', 'EMS', 'CT_Unimed_DN'].

['condición', 'EMS', 'CT_Unimed_DN'].
['osteoporosis', 'EMS', 'CT_Unimed_DN'].

Bigramas:
['sistema cardiovascular', 'EMS', 'CT_Bimed_DN_DA'].
['estudio observacional', 'EMS', 'CT_Bimed_DN_NDA'].
['efecto pleiotrópico', 'EMS', 'CT_Bimed_DN_DA'].
['manera positivo', 'EMS', 'CT_Bimed_NDN_DA'].
['falla cardiaca', 'EMS', 'CT_Bimed_NDN_DA'].
['enfermedad renal', 'EMS', 'CT_Bimed_DN_DA'].

Trigramas:
['niveles de colesterol', 'EMS', 'CT_Trimed_DN_DN'].
['enfermedad de Alzheimer', 'EMS', 'CT_Trimed_DN_NPMED'].

En el caso de los unigramas, de los 9 candidatos, 8 se corresponden con términos del diccionario médico y solo uno, ‘estatinas’, fue marcado como neologismo. Además, se extrajeron seis bigramas, de los cuales ‘sistema cardiovascular’, ‘efecto pleiotrópico’ y ‘enfermedad renal’ estaban conformados por nombres y adjetivos. Por otra parte, se detectaron también bigramas con la combinación nombre de lengua general + adjetivo médico (NDN_DA): ‘manera positivo’ y ‘falla cardiaca’ y ‘manera positiva’; no obstante, solo el primero de ellos se corresponde con un término médico. La combinación nombre médico + adjetivo, arrojó el resultado de ‘estudio observacional’. Por último, para el caso de los trigramas, los dos candidatos pertenecen al dominio médico.

4. Resultados Preliminares

Si bien la creación del diccionario electrónico de Smorph se corresponde con la primera etapa de trabajo del proyecto Fondecyt 11130469, cabe señalar los resultados que se extrajeron mediante la utilización de este fueron adecuados. A continuación, en la tabla 1, se muestran los porcentajes de precisión y cobertura obtenidos hasta el momento.

Tabla 2. Resultados obtenidos

	Cobertura	Precisión
Unigramas	59%	94%
Bigramas	76%	97%
Trigramas	75%	98%

El diccionario médico implantado en Smorph contiene el 59% de los unigramas del corpus y mediante combinaciones se pueden detectar el 76% de los bigramas y el 75% de los trigramas. Asimismo, también se puede observar el alto porcentaje de precisión. Esto demuestra que el método posee una gran efectividad; no obstante, se espera mejorar los resultados, sobre todo en la cobertura, mediante reglas morfológicas, para los neologismos que incluyan formantes cultos; y sintácticas, para los que no posean dichos elementos.

5. Conclusiones

A modo de conclusión, el presente trabajo pretende ser un aporte a las tareas de extracción de información, así como también para los estudios de terminología médica, al presentar el análisis de la

estructura morfológica de los textos y estudiar los contextos sintácticos en los que dichas construcciones aparecen. La propuesta de extracción se asume como una herramienta en la que se pueda observar la manera en que se presentan los términos en contextos reales y de qué manera son utilizados, más que establecer pautas de estandarización y normativización.

De acuerdo con los objetivos planteados en el proyecto, se logró establecer un medio de detección automática de candidatos a término del dominio médico a partir de la utilización de información léxica contenida en diccionarios. Si bien, como se mencionó más arriba, esto no es una solución definitiva, constituye un primer paso en este tipo de tareas. Asimismo, también se comprobó que recurrir a la combinación entre expresiones médicas y neologismos puede constituir una manera adecuada para la detección de nuevos candidatos. No obstante, pueden surgir varios inconvenientes, los cuales son necesarios tener en cuenta. El más importante es que no toda PD es un candidato a término, pues, en los textos que conforman el corpus, se encuentran nombres propios (de autores, países, etcétera), erratas, etcétera, que dificultan las tareas de extracción. Al respecto, una de las soluciones es recurrir a la morfología, puesto que ya en Vivaldi [29], se mencionaba la posibilidad de recurrir a este tipo de pistas en la detección. Por otro lado, en casos de que esto no sea suficiente, en el proyecto se contempla la información que brinda el contexto sintáctico. Al respecto, se realizó una investigación preliminar utilizando la información léxica, morfológica y sintáctica [7], con resultados adecuados. Asimismo, también es pertinente observar que algunos términos pueden presentar estructuras diferentes a las planteadas en el presente trabajo, por lo que, a tales efectos, en etapas posteriores, se pretende desarrollar reglas de detección que consideren otras posibles combinaciones de palabras.

El trabajo a futuro se organiza en torno a los siguientes ejes: (i) elaborar reglas de extracción basadas en información morfológica; (ii) elaborar reglas de extracción basadas en deducción de categorías gramaticales a partir del contexto sintáctico; (iii) elaborar reglas de filtrado para evitar etiquetados erróneos; (iv) extender las tareas de extracción a estructuras diferentes, como ser, por ejemplo, la secuencia ‘nombre adjetivo adjetivo’ (‘accidente isquémico transitorio’) o bien extensiones mayores: cuatrigramas (‘bacilo de la tuberculosis’), quintigramas (‘bacilo de la tuberculosis aviar’), etcétera.

Referencias

1. Diccionario Mosby (versión electrónica). Harcourt. Madrid (2005)
2. Cárdenas, E.: Terminologìa Médica. Mc Graw Hill. México D. F. (2012)
3. Aït-Mokthar, S.: SMORPH: Guide d’utilisation. Rapport technique. Clermont-Fd.: Universidad Blaise Pascal/GRILL (1998)
4. Abacci, F.: Développment du Module Post-Smorph. Tesis (Maestría en informática). Memoria del DEA de Lingüística e Informática, Universidad Blaise-Pascal, Clermont-Ferrand (1999)
5. Koza, W.: Extracción de candidatos a término del dominio médico a partir de la categorización automática de palabras. Congreso Argentino de Informática y Salud CASI 2012, La Plata, Argentina (2012) 153-160
6. Koza, W.; Koza, S.; Muñoz, M.; Ojeda, M. & Yepes, E.: Development of an automatic extractor of medical term candidates with linguistic techniques for Spanish. Seconde International Conference on Informatics and Applications (ICIA) 2013, Lodz, Poland, (2013) 53-58
7. Koza, W. Propuesta de extracción automática de candidatos a término del dominio médico procesando información lingüística. Descripción y evaluación de resultados. Alfa. Revista de Linguística. (2015) 113-127
8. López-Huertas, M. Barité, M. & Torres, I.: Terminological representation of specialized areas in conceptual structures: the case of gender studie. In: López-Huertas, M. Barité, M. & Torres, I. International ISCO Conference 8, 2004, London. Proceedings… London: Ia C. McIlwaine, (2004) 263-268.
9. Sager, J.: Pour une approche fonctionnelle de la terminologie. In: Bènjoint, H; Thoiron, P. (Ed.). Le sens en terminologie. Lyon: Presses Universitaires de Lyon, (2000) 40-60
10. Marincovich, J.: Palabra y término: ¿Diferenciación o complementación?. Revista Signos: Estudios de Lingüística, Valparaíso, v.41, n.67, (2008) 119-126
11. Jacquemin, C. & Borigault, D.: Term extraction and automatic indexing. In: Mitkov, R. (Ed.). The Oxford Handobook of Computational Linguistics. Oxford: Oxford University Press, (2005) 599-615
12. Krauthammer, M. & Nenadic, G. Term identification in the biomedical literature. Journal of Biomedical Informatics, San Diego, v.37, n.6, (2004) 512-526.

13. Cabré, M.: Morfología y terminología. En: Felíu, E. La morfología a debate. Jaén: Universidad de Jaén, (2006) 131-144.
14. López, C.; Tercedor, M., Faber, P. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. Revista E-Salud, Málaga, v.2, n.8, (2006) 228-240.
15. Castro, E.; Iglesias, A.; Martínez, P.; Castaño, L.: Automatic identification of biomedical concepts in Spanish language unstructured clinical texts. In: CASTRO, E. et al. In: AC INTERNATIONAL HEALTH INFORMATICS SYMPOSIUM, 1., 2010, Nueva York. Proceedings, Nueva York: ACM (2010) 751-757.
16. http://www.ihtsdo.org/snomed-ct/
17. http://www.nlm.nih.gov/research/umls/
18. Sánchez, D.; Batet, M. & Valls, A.: Web-based semantic similarity: an evaluation in the biomedical domain. Int. J. Software and Informatics, Beijing, v.4, n.1, (2010) 39-52
19. Garla, V. & Brandt, C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatic 2012, Londres, v.13, n.261, (2012). Disponible en: <http://www.biomedcentral.com/1471-2105/13/261>. Acceso en: 30 nov. 2013.
20. Vivaldi, J. & Rodríguez, H.: Using Wikipedia for term extraction in the biomedical domain: first experiences. Procesamiento de Lenguaje Natural, Jaén, v. 45, (2010) 251-254
21. MORENO-SANDOVAL, A.; CAMPILLOS-LLANOS, L. Desing an annotation of multimedica: a multilingual text corpus of the biomedical domain. Procedia: Social and Behavioral Sciences, Amsterdam, v.95, (2013) 33-39
22. Tuason, O. et al. Biological nomenclature: a source of lexical knowledge and ambiguity. In: PACIFIC SYMPOSIUM OF BIOCOMPUTING, 9., 2004, Oak Ridge. Proceedings Oak Ridge: PSB, (2004) 238-249
23. Segura, I.; Martínez, P.; &Samy, D. Detección de fármacos genéricos en textos biomédicos. Procesamiento del lenguaje natural, Jaén, v.40, (2008) 27-34
24. http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council.page
25. Gálvez, C. Reconocimiento y anotación de nombres de fármacos genéricos en la literatura biomédica. Acimed, v.23, n.4, (2012) 326-345
26. http://www.nooj4nlp.net/pages/nooj.html
27. Bès, G. & Solana, Z.: Análisis morfológico y gramáticas locales. Jornadas Argentinas de Lingüística Informática Modelización e Ingeniería 2004, Rosario, (2004)
28. Burdiles, G.: Descripción de la organización retórica del género caso clínico de la medicina a partir del corpus CCM-2009. 2012. 1Tesis Doctoral – Instituto de Literatura y Ciencias del Lenguaje, Facultad de Filosofía y Educación, Pontificia Universidad Católica de Valparaíso, Valparaíso, (2012)
29. Vivaldi, J.: Elaboración de una aplicación automática de reconocimiento y extracción de información terminológica en textos de dominios restringidos. En Cabre, M. & Feliú, J. La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293) (2002) 229-240