PROPUESTA DE EXTRACCIÓN AUTOMÁTICA DE CANDIDATOS A TÉRMINO DEL DOMINIO MÉDICO PROCESANDO INFORMACIÓN LINGÜÍSTICA. DESCRIPCIÓN Y EVALUACIÓN DE RESULTADOS[1]

Walter KOZA ORELLANA[*]

- RESUMEN: Se presenta la descripción de un método de extracción automática de candidatos a términos del área médica a partir del procesamiento de información lingüística. Para ello, se trabajó con reglas en el nivel léxico, morfológico y sintáctico. En primer lugar, se realizó la detección aplicando un diccionario estándar, el cual asignó a las palabras consideradas términos, la etiqueta MED (MÉDICO). Luego, para las palabras que no estaban contempladas en el diccionario (PNCD), se dedujeron las categorías gramaticales apelando a reglas morfológicas y sintácticas. Posteriormente, se procedió a la conformación de sintagmas nominales que involucraban PNCD y MED, para extraerlos como candidatos a términos del dominio. Se utilizaron los softwares Smorph y Módulo Post Smorph (MPS), que trabajan en bloque, y Xfst. Smoprh realiza el análisis morfológico y MPS trabaja sobre gramáticas locales. Xfst, por su parte, es una herramienta de estados finitos que opera sobre cadenas de caracteres, a las que asigna categorías previamente declaradas. El método se probó en una parte del corpus de casos clínicos compilado por Burdiles (2012), que contenía 217258 palabras, y los resultados arrojaron una precisión de 92,58%, una cobertura de 95,02% y una medida f de 93,78%.

- PALABRAS CLAVE: Terminología médica. Extracción automática. Información lingüística. Candidatos a término.

Introducción

El desarrollo sin precedente de las tecnologías de la comunicación ha permitido, principalmente a partir de Internet, la producción, el acceso y el intercambio de un enorme flujo de información y conocimiento científico a usuarios de todo el mundo. No obstante, 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 (LÓPEZ-HUERTAS, BARITÉ; TORRES, 2004). Al mismo tiempo, también es fundamental desarrollar recursos que regulen y analicen los conceptos de

* PUCV – Pontificia Universidad Católica de Valparaíso. Instituto de Literatura y Ciencias del Lenguaje. Facultad de Filosofía y Educación. Viña del Mar – Valparaíso – Chile. 2530388 – walter.koza@ucv.cl.

1 Proyecto FONDECyT, nº 11130469.

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. Ante este nuevo panorama, las investigaciones en el área de la lingüística computacional han realizado diversos aportes a los sistemas de recuperación de información (VILLAYANDRE, 2010) logrando que los usuarios puedan acceder a los datos de manera más rápida y más precisa. Una de las actividades principales en el desarrollo de dichos sistemas es la detección automática de términos de dominios específicos. Un término es una unidad léxica que designa a un concepto en un campo temático particular (SAGER, 2000; MARINCOVICH, 2008), a la vez, desde la perspectiva de la lingüística de corpus, se puede considerar término al *output* de un proceso terminológico (JACQUEMIN; BORIGAULT, 2005).

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, que organizan y especifican el campo de conocimiento, 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 (KRAUTHAMMER; NENADIĆ, 2004). 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 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, en el presente trabajo, se describe el método desarrollado de extracción de candidatos a términos del domino médico a partir del procesamiento de información lingüística. Este trabajo se enmarca en el ámbito la lingüística informática, por un lado, y, por otro, en las tareas de minería textual.

De acuerdo con Cabré (2006), 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

propósitos, las reglas elaboradas en el método que aquí se presenta estuvieron basadas en estos niveles, y se probaron en una parte del corpus compilado por Burdiles (2012) de casos clínicos.

Para el nivel léxico, la detección fue realizada mediante la aplicación de un diccionario estándar, en este caso, el *Diccionario esencial de la Lengua Española* (2006), que le fue cargado al software analizador, el cual asignó a las palabras consideradas términos la etiqueta MED (MÉDICO); en esta tarea se contó con el asesoramiento de expertos del dominio, quienes señalaron los lemas del diccionario de la RAE que pertenecían al área de la medicina. Para las palabras no contenidas en el diccionario (PNCD), se consideró el siguiente planteo: las PNCD que se pueden identificar como nombres o partes de un sintagma nominal son, en su mayoría, expresiones específicas del dominio médico. Vale aclarar que, en el presente trabajo, se tuvo en cuenta el planteo de Moreno-Sandoval (2009), que establece que, por lo general, los sintagmas nominales se corresponden con los términos. A tales efectos, las tareas de extracción se focalizaron en dichos sintagmas.

Se intentó, entonces, deducir la categoría gramatical de las PNCD mediante reglas de formación de palabras y sintácticas. Posteriormente, se procedió a la conformación de sintagmas nominales que involucraban PNCD y MED, extrayéndolos como candidatos a términos del dominio. Finalmente, se evaluó la precisión, la cobertura y la medida F del método.

El trabajo computacional se realizó con las herramientas Smorph (AÏT MOKTHAR, 1998), Módulo Post Smorph (MPS) (ABACCI, 1999) y XFST (BEESLEY; KARTTUNEN, 2003). 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. Xfst es una herramienta de estados finitos que opera sobre cadenas de caracteres, a las que asigna categorías previamente declaradas, para luego, dar lugar al análisis automático de expresiones; para ello, es necesario elaborar un conjunto de reglas que interactúen determinando combinaciones posibles de categorías.

El trabajo se organiza de la siguiente manera. En la sección 2, se presentan los antecedentes en el área. En la sección 3, se describe la metodología y el trabajo realizado; y, en la sección 4, los resultados obtenidos. Finalmente, en la sección 5, se presentan las conclusiones derivadas de la investigación.

Extracción de términos en el área médica

En el campo médico, Krauthamer y Nenadić (2004) mencionan que las barreras para una extracción de términos exitosa incluyen 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 (2006), 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 et al. (2010), 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 (IHTSDO, 2013). 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 (NLM, 2013), se han realizado estudios de reconocimiento automático de similitud semántica. Entre ellos, se pueden mencionar los llevados a cabo por Sánchez, Batet y Valls (2010), y Garla y Brandt (2012). 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 (2010) 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 (2013) 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 y al público en general, como, por ejemplo, manuales y revistas médicas no especializadas. 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*) (TUASON et al., 2004).

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 (2008), 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) (AMA, 2013), 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 (2012) propone un trabajo similar aunque basado solamente en reglas morfológicas, al igual que Segura, Martínez y Sami (2008), propuestas por USAN, y recurriendo a la herramienta de estados finitos NooJ (2013). De esta manera, la autora logra 99,8% de precisión y 92% de cobertura.

Pues bien, en el método aquí presentado, se emplean los dos enfoques mencionados, es decir, tanto información brindada por diccionarios, en este caso, se optó por un diccionario estándar no especializados, como así también, la deducción de palabras no incluidas en dicho diccionario mediante pistas morfológicas. Además, se recurre también a información brindada por el contexto sintáctico. A continuación se describe el trabajo realizado.

Modelización e implantación en máquina

Para desarrollar el método de detección automática de candidatos a términos del dominio médico, se llevó a cabo la elaboración de un conjunto de reglas lexicográficas, morfológicas y sintácticas que permitan detectar las expresiones propias de dicha área.

La metodología del presente trabajo se basa en dos aspectos fundamentales: (i) la asignación de la etiqueta MED (MÉDICO) a las entradas del diccionario de Smorph con el objeto de reconocer, en los textos, aquellos términos específicos del dominio médico que se encontraban en un diccionario estándar, y (ii) deducir la categoría de las palabras que no se encuentran en el diccionario fuente de Smorph mediante: (a) su estructura morfológica y (b) su contexto sintáctico. Para el primer aspecto, se cotejaron los términos propios del área incluidos en el *Diccionario esencial de la lengua española* (2006) (por ejemplo, 'enfermedad', 'médico',

'cáncer', 'presión baja', etcétera); en esta tarea se contó con el asesoramiento de expertos del dominio, quienes señalaron los lemas del diccionario de la RAE que pertenecían al área de la medicina. Para el segundo, se tomaron en consideración los estudios de formación de palabras generales (VARELA, 2005) y propias de la medicina (DURUSSEL, 2006); la relación entre morfología y terminología (CABRÉ, 2006) y los análisis sobre la conformación de sintagmas (NUEVA…, 2010).

Para el trabajo informático, se recurrió a las herramientas Smorph (AÏT MOKTHAR, 1998), Módulo Post Smorph (MPS) (ABACCI, 1999) y Xfst (BEESLEY; KARTTUNEN, 2003) de Xerox.

Smorph es un analizador y generador textual que, en una única etapa, realiza la delimitación previa de los segmentos textuales a considerar y el análisis morfológico, dando las formas correspondientes a un lema con los valores correspondientes. Este programa es una herramienta declarativa, y 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 se le cambia la información lingüística.

Las fuentes declarativas de Smorph están constituidas por 5 archivos: (i) ascii.txt: contiene los códigos ascii específicos tales como los separadores de oración y de párrafo; (ii) rasgos.txt: incluye etiquetas de rasgos morfológicos a aplicar en el análisis de las cadenas de caracteres con sus posibles valores (ej.: EMS: 'nombre', 'verbo'; Género: 'masculino', 'femenino', etcétera); (iii) term.txt: carga las diferentes terminaciones que cada lema puede presentar en su derivación morfológica (ej.: -o, -a, -os, -as); (iv) entradas.txt: es el listado de lemas y modelos correspondientes de derivación (ej. casar v1), y (v) modelos.txt: define las clases de acuerdo con los parámetros de concatenación regular de cadenas a partir de las entradas y las terminaciones (ej.: modelo v1: raíz + terminaciones de la 1ª conjugación regular + rasgos). Una característica del programa es que se puede asignar categorías por defecto, en este caso, a aquellas palabras de los textos que no están en su diccionario, les asigna automáticamente la etiqueta PD (palabra desconocida). A la vez, también puede clasificar palabras de acuerdo con su terminación, lo que Aït Mokthar (1998) denomina 'terminaciones distinguidas', por ejemplo, todas las palabras en español terminadas en '-ción' son nombres femeninos, con lo cual, no sería necesario cargar los nombres con dicha terminación, puesto que bastaría con indicar esa información en el archivo term.txt.

Por su parte, 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.

Las fuentes declarativas de MPS, a diferencia de Smorph, están constituidas por un único tipo de archivo, rcm.txt, que incluye un listado de reglas que especifican cadenas posibles de lemas con una sintaxis informatizada. Las reglas pueden ser de tres tipos:

1. De reagrupamiento: Determinante + Nombre = Sintagma Nominal
2. De descomposición: Contracción = Preposición + Determinante
3. De correspondencia: Artículo = Determinante

Por último, para el caso de Xfst, la aplicación se presenta como una implementación de autómatas de estados finitos, cuyo objetivo es producir análisis morfológico y generación. Esta herramienta trabaja con archivos fuentes en los que se declara la información lingüística en un editor de textos planos (.txt). Entre las herramientas que utiliza este programa, se encuentran los tokenizadores de estados finitos, que ejecutan la segmentación del texto de acuerdo con la información morfosintáctica almacenada. En este caso, se utilizó esta herramienta para localizar aquellos términos médicos que contenían algún tipo de formante propio de la medicina, como, por ejemplo, '-algia', para el caso de 'neuralgia', 'gastralgia'; '-blasto', para el caso de 'blastocito', 'blastoma', etcétera.

El proceso de reconocimiento de PD y posterior extracción de candidatos a términos se abarcó las siguientes etapas:

- **Etapa I:** Análisis morfológico y reconocimiento de los signos de puntuación por medio de Smorph. Aquí se les asignó a las palabras desconocidas la etiqueta 'PD'.
- **Etapa II:** Modificación del archivo term.txt mediante la asignación de terminaciones distinguidas con su correspondiente clasificación morfológica. Posteriormente, se volvió a pasar el corpus por Smorph a fin de obtener las categorías que se ajusten a dichas terminaciones. También en esta etapa se consideró la posibilidad de que la PD sea un nombre propio o una sigla a partir de si presenta o no caracteres en mayúscula.
- **Etapa III:** Reconocimiento de candidatos a términos a partir de estructuras morfológicas, mediante Xfst. En esta etapa se pasó el corpus por la herramienta Xfst con el propósito de detectar aquellas palabras que contengan en su estructura alguna particularidad con los términos médicos. Para ello, a modo de ejemplo, se declaró en el archivo fuente reglas del tipo: 'necro + letra(s) = término médico' (ejemplo: 'necropsia', 'necrosis'); 'letra(s) + cardio + letra(s) = término médico' (ejemplo: 'microcardiopatía', 'electrocardiograma'). A las palabras reconocidas mediante este método, se les asignó la etiquea CT y se adecuó al formato de salida de Smorph.

- **Etapa IV:** Creación y aplicación de reglas sintácticas que permitan deducir la categoría de las PD. Aquí se hizo hincapié en la estructura del sintagma nominal (SN) (Ej.: Det + PD + Adj = SN/ART+NOM+ADJ).
- **Etapa V:** Extracción de los SN que involucran PD, en calidad de candidatos a términos. Aquí los términos fueron simplificados con la técnica de *stemming* (MANNING; RAGHAVAN; SCHÜTZE, 2009), que consiste en reducir las palabras a sus formas no flexivas y no derivativas.
- **Etapa VI:** Evaluación de las categorizaciones y de los candidatos a términos extraídos mediante las medidas de precisión, cobertura y medida F.

El método propuesto se probó en parte del corpus de casos clínicos, CCCM-2009, compilado por Burdiles (2012). Sobre dichos textos, los expertos elaboraron listas de referencias con los términos que allí se encontraban. En estas se incluyeron conceptos propios de la anatomía, síntomas, compuestos químicos, nombres de enfermedades y todo aquello que los expertos consideraban de uso habitual y específico de la medicina.

A continuación se ejemplifica la extracción realizada con un breve fragmento del corpus en donde se reconoció una serie de términos específicos.

Figura 1 – Fragmento del CCCM-2009 analizado

> Enfermedad de tricocefalosis es la infección por **Trichuris trichiura**, parásito que se ubica en el intestino grueso, que con frecuencia se comporta como comensal, pero puede originar sintomatología cuando está presente en gran número, especialmente en niños con deficiencias nutritivas. (Boletín Chileno de Parasitología, v.54, n.3-4, 1999).

Fonte: apud Burdiles (2012).

Aquí, Smorph etiquetó como candidatos a términos 'enfermedad', 'infección', 'parásito', 'intestino grueso', 'comensal', 'sintomatología' y 'deficiencias nutritivas', debido a que dichos términos se hallaban dentro del diccionario fuente. A su vez, etiquetó como palabras desconocidas 'tricocefalosis', 'Trichuris' y 'trichuria'. La identificación de estas últimas fue realizada en las etapas posteriores, aquí explicitadas:

1. Se pasó el texto por xfst, en donde el archivo con las reglas de nivel morfológico contenía la siguiente:

   letra ≥ 1 + cefal + osis + letra ≥ 1 = 'CT'

   Vale aclarar que las expresiones 'cefal' y 'osis' estaban contenidas en la lista de raíces médicas.

2. Se pasó el texto por MPS, en donde el archivo rcm.txt de reglas sintácticas incluía las siguientes:
   Preposición + PD + PD + Signo de puntuación = Prep_SNMED_SigP
   CT + preposición 'de' + CT = Trigrama

3. En el caso de las expresiones etiquetadas como Prep_SNMED_SigP, se eliminó la preposición y el signo de puntuación obteniendo el bigrama 'Trichuris trichuria'.

Como se mencionó, el método propuesto fue evaluado mediante las medidas de precisión, cobertura y medida f. En la sección siguiente, se presentan los resultados obtenidos.

Evaluación de resultados

Los resultados de los experimentos fueron evaluados mediante las medidas de precisión, cobertura y medida f. Los expertos del dominio elaboraron una lista de referencia con un total de 10092 términos distribuidos de la siguiente manera:

- Unigramas: 2367
- Bigramas: 5084
- Trigramas: 2641

Del total de la lista, se reconocieron 9590 y se marcaron erróneamente 769, lo que implicó una precisión de 92,58%, una cobertura de 95,02% y una medida f de 93,78%. A continuación se presenta una tabla en la que están discriminados los resultados en unigramas, bigramas y trigramas.

Tabla 1 – Resultados obtenidos

	Unigramas	**Bigramas**	**Trigramas**
Precisión	79,65%	96,96%	99,25%
Cobertura	97,08%	91,48%	96,02%
Medida F	87,50%	94,14%	97,61%

Fonte: Elaboración propia.

Como se puede apreciar, la mejor precisión la obtuvieron los trigramas, mientras que la mejor cobertura se logró para los unigramas; asimismo, la medida F más adecuada se dio en el caso de los trigramas.

Se detectaron algunos inconvenientes en la precisión de los unigramas, una de las causas fueron algunas palabras comunes que tenían algunos elementos en común con los términos, como por ejemplo 'fotografía'. Para el caso de la cobertura, los problemas se derivaron de palabras médicas no consideradas como tales en el diccionario de la RAE, por ejemplo 'diámetro'. Otro de los problemas fueron los errores de ortografía cometidos por los autores del texto.

No obstante, de acuerdo con los resultados obtenidos, puede considerarse válido el método propuesto.

Consideraciones finales y trabajos futuros

Se presentó un método de detección automática de candidatos a términos del dominio médico mediante la aplicación de técnicas lingüísticas. Se trabajó con reglas en el nivel lexicográfico, morfológico y sintáctico, recurriendo a los programas Smorph, Módulo Post Smorph (MPS) y Xfst.

El método propuesto fue probado en una parte del corpus de casos clínicos, CCCM-2009 compilado por Burdiles (2012), logrando un 95,02% de cobertura, un 92,58% de precisión y una medida F de 93,78%. Los resultados obtenidos permiten suponer que se está ante un método a grandes rasgos efectivo y que abre nuevas perspectivas en torno a la extracción automática de candidatos a términos.

Una cuestión a destacar es que se optó por un diccionario estándar a fin de probar la efectividad de las reglas del orden morfológico y sintáctico. A partir de los resultados obtenidos, se pudo apreciar que, aproximadamente, el 50% de los términos no hallados en el *Diccionario esencial de la lengua española* (DICCIONARIO…, 2006) fueron detectados mediante dichas reglas. No obstante, en una experimentación futura, se trabajará con un diccionario del dominio, *Diccionario de términos médicos* (2012), de la Real Academia Nacional de Medicina y se compararán los resultados.

Las detecciones erróneas se debieron, principalmente, a PD que no presentaban una estructura morfológica propia de la medicina y, a la vez, se hallaban aisladas o los demás elementos que las rodeaban no eran suficientes para deducir su categoría gramatical, por ejemplo en una lista vertical o entre paréntesis. Por otro lado, también hay que señalar los casos de nombres propios que, en algunas ocasiones, pueden ser términos, como por ejemplo 'Alzheimer', lo que implica que no se pueden descartar desde un primer momento. Por último, hay que mencionar que los errores de ortografía y de tipiado que presentaban algunos textos.

La ventaja principal en este tipo de métodos es que puede demostrar su efectividad no solo en grandes masas textuales, sino también en corpus más pequeños, con menor cantidad de palabras, se supone que esto ayudaría a las tareas de clasificación automática de documentos a partir de los términos extraídos.

El presente trabajo pretende ser un aporte a las tareas de extracción de información, como así 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.

El trabajo a futuro se organiza en torno a los siguientes ejes. En primer lugar, se pretende adicionar información léxica específica del *Diccionario de términos médicos* (2012). En segundo lugar, se intentará adicionar al método propuesto técnicas de nivel estadístico. En tercer lugar, se realizará el análisis y desarrollo de reglas para la captura automática de la variación denominativa. Finalmente, en cuarto lugar, se considerarán posibles técnicas de clasificación automática de documentos a partir de los términos extraídos con el presente método.

KOZA ORELLANA, Walter. Proposal for an automatic extraction for medical term candidates processing linguistic information. Description and evaluation of results. **Alfa**, São Paulo, v.59, n.1, p.113-127, 2015.

- *ABSTRACT: The description of a method for automatic extraction of term candidates from the medical field by applying linguistic information is presented. Lexicography, morphological and syntactic rules were used. First, the detection was performed by applying a standard dictionary that assigned the tag 'MED' ('MEDICAL') to the words that could be considered terms. Morphological and syntactic rules were used to try to deduce the part of speech of the words that were not considered in the dictionary (WNCD). Afterwards, nominal phrases that included WNCD and MED were gathered to extract them as term candidates of the field. Smorph, Post Smorph Module (MPS) – both working in groups – and Xfst were the software used. Smorph performs the morphological analysis of character strings and MPS works on local grammar. Xfst is a finite state tool that works on character strings assigning previously stated categories to allow the automatic analysis of expressions. This method was tested on a section of the corpus of clinical cases collected by Burdiles (CCCM - 2009) containing 217,258 words. The results showed 92.58% of precision, 95.02% of recall and 93.78% of F-measure.*

- *KEYWORDS: Medical terminology. Automatic extraction. Linguistic information. Terms candidate.*

REFERENCIAS

ABACCI, F. **Développement du module post-smorph.** 1999. Tesis (Maestría en Informática) – Memoria del DEA de Lingüística e Informática, Universidad Blaise-Pascal, Clermont-Ferrand, 1999.

AÏT MOKTHAR, S. **SMORPH:** guide d'utilisation: rapport technique. Clermont: Universidad Blaise Pascal: GRILL, 1998.

AMERICAN MEDICAL ASSOCIATION [AMA]. **United States adopted names council.** Disponible em: <http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council.page>. Acceso en: 15 nov. 2013.

BEESLEY, K.; KARTTUNEN, L. **Finite state morphology.** Stanford: CSLI Stanford University, 2003.

BURDILES, G. **Descripción de la organización retórica del género caso clínico de la medicina a partir del corpus CCCM-2009.** 2012. 199p. Tesis Doctoral – Instituto de Literatura y Ciencias del Lenguaje, Facultad de Filosofía y Educación, Pontificia Universidad Católica de Valparaíso, Valparaíso, 2012.

CABRÉ, M. Morfología y terminología. In: FELÍU, E. **La morfología a debate.** Jaén: Universidad de Jaén, 2006. p.131-144.

CASTRO, E. et al. Automatic identification of biomedical concepts in Spanish language unstructured clinical texts. In: CASTRO, E. et al. In: ACM INTERNATIONAL HEALTH INFORMATICS SYMPOSIUM, 1., 2010, Nueva York. **Proceedings…** Nueva York: ACM, 2010. p.751-757.

DICCIONARIO esencial de la lengua española. Madrid: RAE, 2006.

DICCIONARIOS de términos médicos. Buenos Aires: Editorial Médica Panamericana, 2012.

DURUSSEL, B. **Terminología médica.** Santa Fe: Universidad Nacional del Litoral, 2006.

GÁLVEZ, C. Reconocimiento y anotación de nombres de fármacos genéricos en la literatura biomédica. **Acimed**, La Habana, v.23, n.4, p.326-345, 2012.

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.

IHTSDO. **SNOMED:** The global language of healthcare. Disponible en: <http://www.ihtsdo.org/snomed-ct/>. Acceso en: 15 nov. 2013.

JACQUEMIN, C.; BORIGAULT, D. Term extraction and automatic indexing. In: MITKOV, R. (Ed.). **The Oxford Handobook of Computational Linguistics.** Oxford: Oxford University Press, 2005. p.599-615.

KRAUTHAMMER, M.; NENADIĆ, G. Term identification in the biomedical literature. **Journal of Biomedical Informatics,** San Diego, v.37, n.6, 512-526, 2004.

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, p.228-240, 2006.

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. p.263-268.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval.** Cambridge: Cambridge University Press, 2009.

MARINCOVICH, J. Palabra y término: ¿Diferenciación o complementación?. **Revista Signos:** Estudios de Lingüística, Valparaíso, v.41, n.67, p.119-126, 2008.

MORENO-SANDOVAL, A. Panorama actual de la ingeniería lingüística. In: AMPARO, A.; RAMBLA, E.; VALERO, E. (Ed.). **Terminología y sociedad del conocimiento.** Berlín: Peter Lang Bern, 2009. p.99-116.

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, p.33-39, 2013.

NATIONAL LIBRARY OF MEDICINE [NLM]. **Unified Medical Language System (UMLS).** Disponible en: <http://www.nlm.nih.gov/research/umls/>. Acceso en: 15 nov. 2013.

NOOJ. Disponible en: <http://www.nooj4nlp.net/pages/nooj.html>. Acceso em: 15 nov. 2013.

NUEVA gramática de la lengua española. Madrid: RAE, 2010.

SAGER, J. Pour une approche fonctionnelle de la terminologie. In: BÉJOINT, H.; THOIRON, P. (Ed.). **Le sens en terminologie.** Lyon: Presses Universitaires de Lyon, 2000. p.40-60.

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, p.39-52, 2010.

SEGURA, I.; MARTÍNEZ, P.; SAMY, D. Detección de fármacos genéricos en textos biomédicos. **Procesamiento del lenguaje natural,** Jaén, v.40, p.27-34, 2008.

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. p.238-249.

VARELA, S. **Morfología lexica:** la formación de palabras. Madrid: Gredos, 2005.

VILLAYANDRE, M. **Aproximación a la lingüística computacional.** León: Universidad de León, 2010.

VIVALDI, J.; RODRÍGUEZ, H. Using Wikipedia for term extraction in the biomedical domain: first experiences. **Procesamiento del Lenguaje Natural,** Jaén, v.45, p.251-254, 2010.

Recebido em dezembro de 2013

Aprovado em fevereiro de 2014