Extracción terminológica en el dominio médico a partir del reconocimiento de sintagmas nominales [0]

Term extraction in the medical domain from noun phrases recognition

Walter Koza, Zulema Solana
Universidad Nacional de Rosario-CONICET-INFOSUR
Rosario, Argentina
kozawalter@opendeusto.es, zsolana@arnet.com.ar

Merley da S. Conrado, Solange O. Rezende, Thiago A. S. Pardo
Universidade de São Paulo
São Paulo, Brasil
merleyc@icmc.usp.br, solange@icmc.usp.br, taspardo@icmc.usp.br

Josuka Díaz-Labrador, Joseba Abaitua
Universidad de Deusto
Bilbao, España
josuka@deusto.es, joseba.abaitua@deusto.es

Abstract

The tasks of term extraction have a special place in activities of extraction and organization of knowledge. Aiming at comparing different proposals for this task, in this paper we describe two approaches to the automatic extraction of medical terms through noun phrases (NPs) recognition on medical text corpus in Spanish. In the first approach, we extracted all NPs that were considered as our *baseline*. In the second one, the extraction process used specific NPs, which were determined on the basis of syntactic and positional criteria, among others.

We worked on the IULA corpus [1] constituted by medical texts in Spanish and results were compared to reference lists also provided by IULA. The tools Smorph [2] and MPS [3] were used. From the experiments, we were able to obtain three interesting contributions to the extraction tasks: (i) we showed that it is possible to extract medical terms from specific NPs recognition, (ii) the software dictionary used was improved with 2,445 new terms, and (iii) terms that were not in the reference lists were extracted. For the third contribution, we used the SNOMED CT® [4] terms lists, aiming at improving the IULA reference lists.

Keywords: Term extraction, Noun Phrase, Medical Terminology, Automatic Recognition, SNOMED CT®

Resumen

Las tareas de extracción de términos tienen un lugar destacado en las actividades de extracción y organización del conocimiento. Con el objetivo de comparar diferentes propuestas para esta tarea, en el presente artículo se describen dos aproximaciones para la extracción automática de términos médicos a partir del reconocimiento automático de sintagmas nominales (SN) realizado sobre un

corpus de textos médicos en español. En el primero de ellos, considerado como *baseline*, se extrajeron todos los SN encontrados. En el segundo experimento, en cambio, la extracción estuvo dirigida a SN específicos, que fueron determinados sobre la base de criterios sintácticos y posicionales, entre otros.

Trabajamos con el corpus IULA [1] compuesto por textos de medicina en español y los resultados fueron comparados con las listas de referencias de dicho corpus. Se trabajó con las herramientas de análisis lingüístico Smorph [2] y Módulo Post Smorph (MPS) [3]. A partir de los experimentos, se logró realizar tres contribuciones a las tareas de extracción: (i) se demostró que es posible extraer términos médicos mediante el reconocimiento de sintagmas nominales específicos, (ii) se le adicionaron 2445 nuevos términos al diccionario fuente de Smorph y (iii) se extrajeron términos que no estaban en la lista de referencia. Para la tercera contribución, se utilizó la lista de términos de SNOMED CT® [4] con el objetivo de mejorar las listas de referencias de IULA.

Palabras Claves: Extracción de términos, Sintagma nominal, Terminología Médica, Reconocimiento Automático, SNOMED CT®.

1. INTRODUCCIÓN

Las tareas de extracción de términos poseen un lugar destacado en actividades de extracción y organización del conocimiento. Un término es una unidad léxica caracterizada por una referencia especial dentro de una disciplina [5], y puede estar conformado por una sola palabra (unigrama), por ejemplo ‘asma’, ‘hormona’; o una combinación de ellas (n-gramas), como ‘tuberculosis pulmonar’, ‘sistema cardiovascular’ (bigramas); ‘esquema de tratamiento’, ‘estado de enfermedad’ (trigramas), etcétera. Un conjunto de términos constituye la terminología.

Este tipo de tareas suele enfocarse en dominios específicos, y uno de ellos es el de la medicina. En este caso, la extracción de términos representativos suele destinarse a la elaboración de listas de entradas para diccionarios electrónicos específicos, la creación de base de datos o de ontologías y taxonomías, que organizan y especifican el dominio de conocimiento, etcétera. Otra de las aplicaciones apunta a la clasificación textual, es decir que, a partir de la extracción de términos realizada sobre varios textos o informes de medicina, es posible reconocer a qué subáreas pertenecen dichos textos.

Existen diversos trabajos en la literatura que realizan la extracción de términos [6,7,8,9]. En relación con el dominio de medicina, de acuerdo con Castro y sus colaboradores [10], para el caso del inglés, hay varias investigaciones orientadas al procesamiento de textos y de datos de ese dominio [11,12], sin embargo, se encuentran pocas iniciativas para el español [13,14,15,16,17].

A modo de aporte, en el presente trabajo se describen dos extracciones de términos realizadas a partir de la identificación de sintagmas nominales en un corpus de textos médicos en español. El objetivo fue comparar extracciones hechas de diferentes modos. En la primera, se consideraron candidatos a término todos los sintagmas nominales (SN), siendo esta nuestro *baseline*; y en la segunda, las tareas de extracción estuvieron dirigidas a SN específicos, que se eligieron sobre la base de criterios sintácticos, posicionales, entre otros. Para los experimentos se utilizó el corpus IULA [1] de medicina en español y los resultados se compararon con tres listas de referencia correspondientes a unigramas, bigramas y trigramas, lo que permitió arribar a dos cuestiones interesantes: (i) demostrar que la identificación de sintagmas específicos permite extraer términos representativos del dominio de medicina y (ii) señalar que algunos términos extraídos que no

estaban en las listas de referencia podían llegar a ser términos importantes del dominio. Para el segundo de los objetivos, se utilizo una lista de términos consagrada, el SNOMED CT® [4], con el propósito de mejorar las listas de referencia.

El artículo se organiza de esta manera: en el parágrafo siguiente, se presenta un breve estado de la cuestión referente a la extracción de términos en medicina; en 3, se mencionan las herramientas informáticas para el reconocimiento de sintagmas; en 4, se presenta la metodología para la extracción de términos; en 5, la descripción de los experimentos realizados, y, finalmente, en 6, las conclusiones derivadas de la investigación.

2. Extracción de términos en medicina

Krauthammer y Nenadic [18] elaboran una descripción detallada de los pasos a seguir en la extracción de términos en el terreno de la biomedicina, a la vez que dan cuenta de los experimentos se han estado realizando. De acuerdo con ellos, los términos (nombres de genes, proteínas, drogas, etcétera, para el caso de la biomedicina) son los medios a los que recurre la comunidad científica para identificar conceptos del dominio. Se han venido desarrollando diversos sistemas de reconocimiento automático de términos (RAT) para muchas clases de entidades biomédicas, en particular, para nombres de genes y proteínas. Dichos sistemas se basan tanto en características internas de las palabras correspondientes a clases específicas o en "pistas externas" que pueden ayudar al reconocimiento de secuencias de palabras que representan conceptos del dominio específico. Se toman en consideración diferentes cuestiones, tales como ortografía (mayúsculas, dígitos, caracteres griegos) y "pistas morfológicas" (afijos específicos, etiquetas POS) o información sintáctica proveniente del análisis de superficie. Además, se sugieren diferentes medidas estadísticas para "promover" candidatos a términos, a términos.

Para estas tareas se han desarrollado diversos métodos basado en distintos enfoques, los principales son los siguientes:

(i) Enfoques basados en diccionarios

Los métodos basados en diccionarios para los RAT usan recursos terminológicos existentes a fin de localizar ocurrencias de términos en textos. El problema de este enfoque es que muchas ocurrencias pueden no ser reconocidas si se recurre a diccionarios o base de datos estándares. Por otro lado, también influyen negativamente factores como la homonimia (términos que comparten sus representaciones léxicas con "palabras comunes") y las variaciones en el deletreado de los términos, que incluirían variaciones de "puntuación" (bmp-4 y bmp4), uso de diferentes numerales (syt4 y syt iv) o diferentes transcripciones de las letras del alfabeto griego (iga e ig alpha), o variaciones en el orden de palabras (integrin alpha 4, alpha4 integrin).

(ii) Enfoques basados en reglas

Los enfoques basados en reglas, que es el que nos concierne, intentan recuperar términos por el restablecimiento asociado a los patrones de formación de términos que han sido utilizados para construir los términos en cuestión. El principal enfoque es (en general, manualmente) desarrollar reglas que describan las estructuras de denominación común para ciertas clases de términos usando tanto pistas ortográficas o léxicas, o características morfosintácticas más complejas.

También, en muchos casos, se usan diccionarios de constituyentes de términos típicos (por ejemplo, núcleos terminológicos, afijos, acrónimos específicos) para asistir en el reconocimiento de términos. Sin embargo, los enfoques de la ingeniería del conocimiento son conocidos por ser extremadamente lentos para el desarrollo y, al contar generalmente con reglas muy específicas, sus adaptaciones a otras entidades son usualmente dificultosas.

(iii) Enfoques basados en maquinas de aprendizaje y estadística

Se utiliza una gran variedad de maquinas de aprendizaje (MA) y técnicas estadísticas para los RAT. Mientras los enfoques estadísticos principalmente abordan el reconocimiento de términos generales, los sistemas de MA son usualmente designados para clases específicas de entidades y, así, integran el reconocimiento y la clasificación de términos. Los sistemas de MA usan datos entrenados para "aprender" características útiles para el reconocimiento y la clasificación de términos, pero la existencia de recursos de formación confiables es uno de los principales problemas, ya que no están ampliamente disponibles.

(iv) Enfoques híbridos

Muchos enfoques combinan diferentes métodos (por lo general, reglas y bases estadísticas) y varios recursos (listas precompiladas de términos específicos, palabras, afijos, etcétera) para las tareas de reconocimiento de términos.

En lo que atañe a trabajos en el español, pueden mencionarse los aportes del proyecto ONCOTERM: Sistema bilingüe de información y recursos oncológicos; el sistema Describe®; los trabajos de Vivaldi y Rodríguez y Vivaldi et al.; el trabajo de Castro y sus colaboradores, y la extensa terminología desarrollada por el proyecto SNOMED CT®.

ONCOTERM [13,14] es un proyecto de investigación interdisciplinar sobre terminología médica llevado a cabo por la Universidad de Granada y el Hospital Universitario Virgen de las Nieves. Tiene por objetivo ofrecer, a los implicados (pacientes, profesionales de la medicina, documentalistas de hospital, etcétera), la información disponible relacionada con el cáncer, tanto en español como en inglés. A tales efectos, han trabajado en el desarrollo de corpus de textos médicos pertenecientes al cáncer, en bases de datos terminológicas basadas en el conocimiento, donde se incluyen hipervínculos con base textual y en la integración de esos recursos en un servidor web. El trabajo terminográfico de este grupo posee un enfoque descriptivo con la intención de dar cuenta de la competencia lingüística del experto y sigue el Modelo Lexemático-Funcional (MLF) diseñado por Mingorance [19], que constituye una aproximación onomasiológica al estudio de la representación del significado [20].

El sistema Describe® [21], por su parte, aplica un Extractor de Contextos Definitorios [22] para la búsqueda, clasificación y agrupamiento de definiciones en la Web. Este sistema utiliza robots para indexar de manera constante páginas que contengan alguno de los dos millones de términos relacionados con el área de medicina. Dichas páginas constituyen la base de datos inicial para la extracción de contextos definitorios, una vez extraídos los diferentes tipos de definiciones, estos se clasifican según su tipo y se agrupan de acuerdo con el contenido semántico que en ellos se vincula [23].

Vivaldi y Rodríguez [15] toman como base de información semántica a la Wikipedia para crear un sistema de extracción de términos. La metodología consiste en tomar un documento y su correspondiente conjunto de candidatos a términos para comparar los resultados que se obtienen (i) el Coeficiente de Dominio y un conjunto de Marcadores de Dominio (MD), como el definido por YATE [24] (y posteriormente utilizando EuroWordNet [25]); y (ii) una aproximación similar utilizando Wikipedia en lugar de EuroWordNet. Para ello, se usó un MD simple que se correspondiera con la categoría de Wikipedia que remitiese al nombre "medicina". El sistema se probó en un corpus médico en español y luego de comparar los resultados, concluyen que la Wikipedia es un recurso válido para utilizar en esta tarea.

Vivaldi et al [16] utilizan la extracción de términos en un nuevo algoritmo de resumen automático de textos de medicina en español. Para la extracción, los autores utilizan YATE [24], la primera herramienta utilizada para identificar términos médicos a partir de artículos. La extracción permitió

la realización de resumen de textos, que es el foco de artículo, a partir de 50 artículos médicos en español. Los autores afirman que los resultados fueron buenos pero que podrían haber sido mejores si se utilizaba SNOMED CT® [4] para la extracción de términos.

El trabajo de Castro et al [10] presenta una anotación semántica de las notas clínicas en español y la aplicación de una herramienta automatizada para identificar conceptos biomédicos en la ontología de SNOMED CT® en español. Además, los autores presentaron una evaluación de la herramienta utilizando 100 notas clínicas en español anotadas manualmente y afirmaron tener buenos resultados para el dominio de medicina.

El SNOMED CT® [4] (*Systematized Nomenclature of Medicine Clinical Terms*) es una amplia terminología clínica en español, producto de la fusión entre SNOMED RT y *Clinical Terms Version 3*, una terminología previamente conocida como *Codigos Read*, creada en nombre del National Health Service (NHS) (Departamento de Salud del Reino Unido) y propiedad de la Corona Britanica.

3. Sobre las herramientas para el reconocimiento de sintagmas

En el presente trabajo, se adoptó la posición de Moreno [26], que determina que, por lo general, son los sintagmas nominales los que se corresponden con los términos, sean estos conformados por un nombre únicamente o un nombre más complementos. A tales efectos, la extracción automática de candidatos a términos realizada se basó en la detección de este tipo de sintagmas.

Para el presente trabajo, se utilizaron las herramientas SMORPH [2] y Módulo Post-SMORPH (MPS) [3]. SMORPH es un analizador y generador textual que en una única etapa realiza la delimitación previa de los segmentos textuales a considerar (tokenización) y el análisis morfológico (lematización) dando como resultado 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 [2]. MPS, por su parte, ha sido especificado en el GRIL por Caroline Hagège, José Rodrigo, Gabriel Bès y Faizza Abacci, e implantado en C++ en un contexto de Windows por Faiza Abacci [3]. Posteriormente, fue extendido en Pasmo, en donde se le adicionaron otras funcionalidades. MPS, por su parte, toma como input el output de Smorph y, a partir de reglas de (i) reagrupamiento: Det + N = SN; (ii) descomposición: Contracc = P + D, y (iii) de correspondencia: Art = Det.

4. Metodología para Extracción de Términos

Para el presente trabajo, se utilizaron los siguientes sintagmas: sintagmas nominales (SN), sintagmas preposicionales (SP) y sintagmas verbales núcleos (svn). El sintagma núcleo es un bloque casi inseparable que permite reconocer su inicio y su fin, y a partir de propiedad de linealidad se restringen sus posibilidades combinatorias. Esto hace que no presente demasiados problemas de ambigüedad [27]. Se siguió la metodología descrita en la Figura 1.

La extracción comienza con la delimitación del dominio y del corpus con el que se va a trabajar. Inmediatamente después se hace una normalización ortográfica en la que es necesario cambiar la codificación de los archivos del corpus utilizado para UTF-8. En esa normalización, también se sacan manualmente los cambios de líneas para que la herramienta de análisis morfológico no tenga problemas en su procesamiento. Posteriormente se desarrollan reglas para realizar de manera automática la tokenización y el análisis morfológico.

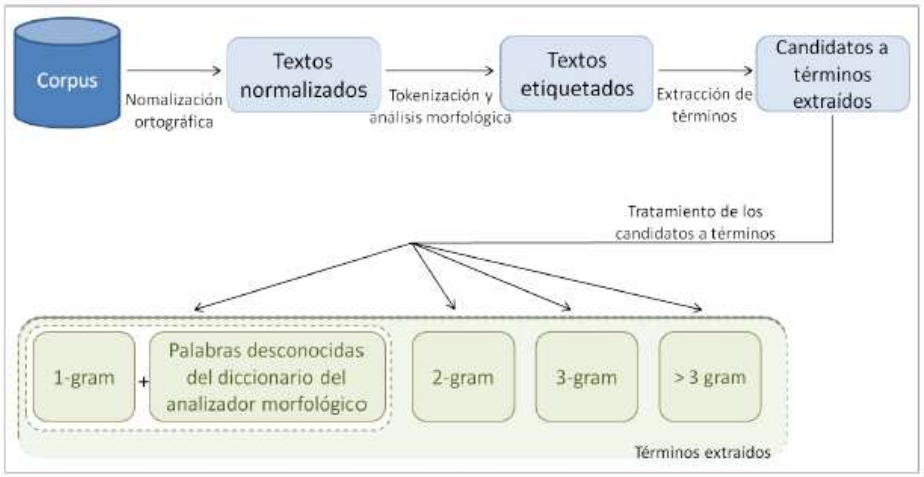

Figura 1. Metodología para la extracción de términos

Es importante destacar que puede haber palabras que no estén en el diccionario fuente de la herramienta, lo que puede afectar el análisis morfológico. A tales efectos, en esta ocasión, aquellos términos no reconocidos en el inicio, se cargaron en el diccionario fuente y se volvió a pasar el corpus por el programa. La existencia de esas etiquetas permite, luego, la constitución de sintagmas y la posterior extracción de aquellos sintagmas de interés del dominio, es decir, de los candidatos a términos.

Asimismo, es necesario hacer un tratamiento de los candidatos a términos con el objetivo de **mantener** solamente los buenos candidatos. Para eso, se remueven las palabras que no añaden a la **representatividad** del dominio y las stopwords. Se denomina stopword (o "palabra vacía") a aquellas palabras que no acrecientan representatividad a los términos (por ejemplo, el artículo "el" en "el asma") o que por sí solas no constituyen términos, como los adverbios, preposiciones, artículos y pronombres. Después de tratar los candidatos a términos, estos se separan en listas de unigramas, bigramas, trigramas y mayores que trigramas para permitir su evaluación. Las palabras que no fueron reconocidas por el analizador morfológico en su primera ejecución, fueron añadidas en la lista de unigramas después de una evaluación manual.

5. Experimentos realizados

5.1. Descripción de los experimentos

Para los experimentos realizados, se utilizó una colección de textos en español con listas de términos de referencia del mismo dominio, el Corpus Técnico IULA-UPF [1]. Dicho corpus es el resultado final de un proyecto de investigación del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, que recopila textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de economía, derecho, medio ambiente, medicina, genómica e informática. Para la elaboración del presente trabajo, utilizamos el corpus de medicina, que contiene un total de 99480 palabras. Junto con él, el IULA-UPF también nos ha facilitado tres listas de términos de referencia, conteniendo unigramas, bigramas y trigramas respectivamente.

La disponibilidad de las listas de términos de referencia fue el principal motivo por el que hemos utilizado ese corpus, ya que estas eran necesarias para evaluar los resultados de los experimentos. Los unigramas de la lista de referencia estaban conformados por un nombre ("alergia") y sumaban un total de 697 unigramas; los bigramas, por un nombre más un adjetivo ("acido benzoico") y sumaban un total de 665 bigramas, los trigramas, por un nombre más la preposición "de" más otro nombre ("grupo de riesgo") y sumaban un total de 82 trigramas.

Las tareas de extracción se realizaron a partir del reconocimiento de SN específicos de acuerdo con distintas subclasificaciones propuestas (ver Figura 2).

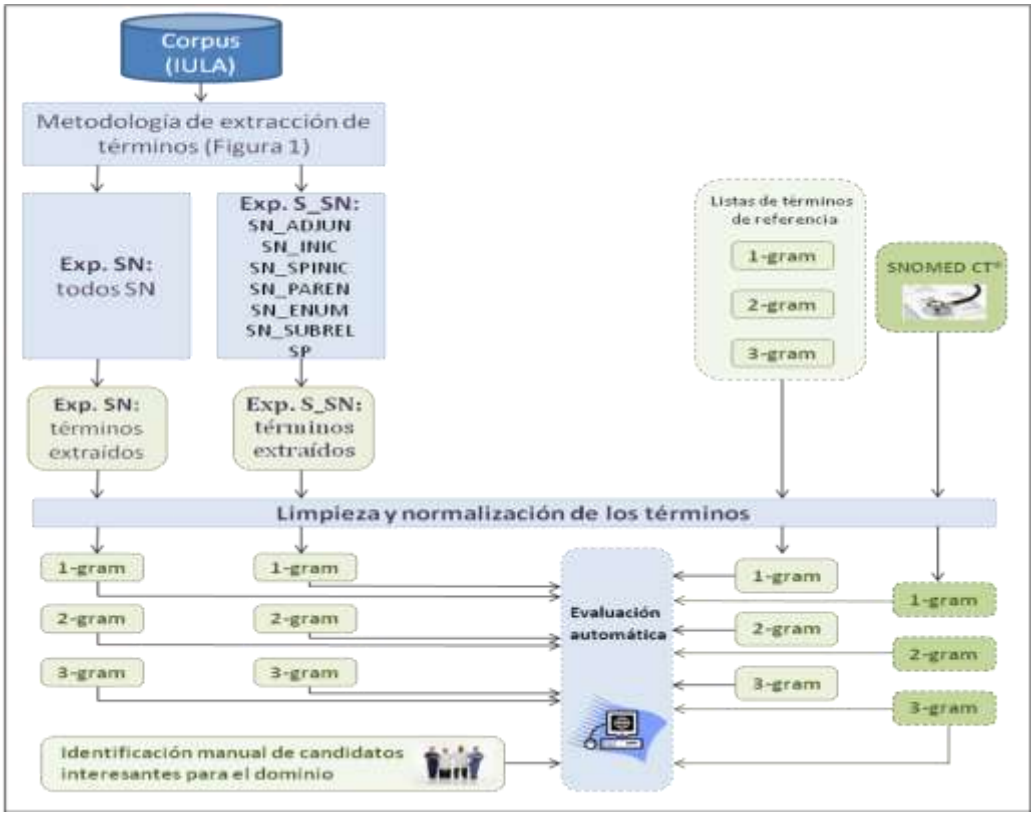

Figura 2. Experimento realizado

A partir del corpus, fue aplicada la metodología de extracción de términos detallada en la Sección 4 (Figura 1). Se realizaron dos experimentos diferentes considerando la subclasificación de los sintagmas nominales. Para ambos, se hizo el análisis morfológico de las palabras que conformaban los textos del corpus mediante el programa SMORPH [2]. A modo de ejemplo, para el fragmento "Pruebas de provocación bronquial con ejercicio y con histamina en niños asmáticos.", el resultado del análisis sería el que se muestra en la Tabla 1.

Tabla 1. Análisis de Smorph [28]

'**Pruebas**' 'prueba', 'EMS', 'nom', 'GEN', 'fem', 'NUM', 'pl']. 'probar', 'EMS', 'v', 'EMS', 'ind', 'PERS', '2a', 'NUM', 'sg', 'TPO', 'pres', 'TR', 'irr', 'TC', 'c1', 'TDIAL', 'est']. '**de**'. ['de', 'EMS', 'prde']. '**provocación**'. ['provocación', 'EMS', 'nom', 'GEN', 'fem', 'NUM', 'sg']. '**bronquial**'. ['bronquial', 'EMS', 'adj', 'GEN', '_', 'NUM', 'sg']. '**con**'. ['con', 'EMS', 'prep'].	'**ejercicio**' ['ejercicio', 'EMS', 'nom', 'GEN', 'masc', 'NUM', 'sg']. '**y**'. ['y', 'EMS', 'cop']. '**con**'. ['con', 'EMS', 'prep']. '**histamina**'. ['histamina', 'EMS', 'nom', 'GEN', 'fem', 'NUM', 'sg']. '**en**'. ['en', 'EMS', 'prep']. '**niños**'. ['niño', 'EMS', 'nom', 'GEN', 'masc', 'NUM', 'pl']. '**asmáticos**'. ['asmático', 'EMS', 'adj', 'GEN', 'masc', 'NUM', 'pl']. '.' ['linsig', 'EMS', 'pun'].

Había un total de 2.445 palabras que no fueron identificadas por el analizador SMORPH y fueron analizadas y añadidas posteriormente al diccionario fuente con el que cuenta el programa.

Una vez obtenidos los resultados de SMORPH, se desarrollaron reglas de reconocimiento de sintagmas. Esas reglas se cargaron en el programa MPS que toma como entrada la salida de SMORPH.

Se realizaron dos experimentos distintos, en los cuales varió la subclasificación de los SN. En el primer experimento (Exp. SN), se consideraron candidatos a término todas las expresiones etiquetadas previamente como SN. En el segundo (Exp. S SN), después de observaciones manuales sobre los términos, se subclasificaron a algunos SN que pudieran tener cierta relevancia, por ejemplo, a partir de cuestiones de orden sintáctico. Dicha subclasificación consideró la posibilidad de que el SN:

- Fuese un adjunto verbal (SN ADJUN): “asoció bronconeumonía”. Para ello, se creó la regla correspondiente con la siguiente estructura svn + SN = SN ADJUN;
- Fuese el antecedente de una cláusula subordinada relativa explicativa (SN SUBREL): “el asma, que se traduce...”. Aquí se requirieron varias reglas, un ejemplo de ellas sería SN + coma + relativo + svn = SN SUBREL. Se crearon reglas para la detección de cláusulas subordinadas relativas explicativas desde el inicio, en el relativo, hasta el verbo.
- Fuese un elemento de una enumeración (SN ENUM): “broncolabilidad, broncorreactividad y reflejos tusígenos broncoconstrictores". Un ejemplo de regla de enumeración sería: SN + coma + SN + conjunción copulativa + SN = ENUM NOM COMP (Enumeración Nominal Completa);
- Apareciera entre paréntesis (SN PARENT): “(fenoterol)”. Se constituyó la regla correspondiente con la estructura paréntesis + SN + paréntesis = SN PARENT;
- Apareciera al inicio de la cláusula (SN INIC): “. Mecanismo inmunológico”. En este caso, para la confección de la regla, se tomó en consideración al punto de la oración anterior: punto + SN = SN INIC. Se decidió tomar como término al SN que apareciera al inicio de la cláusula, puesto que se considera que en esta posición podría cumplir la función de sujeto o ser un elemento topicalizado y previamente se había determinado que ambos fenómenos eran relevantes al momento de la extracción terminológica;
- Fuera adjunto de un sintagma preposicional (SP) ubicado al inicio de la cláusula (SN SPINIC). Al igual que con el SN INIC, se consideró que el SN SPINIC fuera un elemento topicalizado. Por: “. Con salbutamol”. Igual que en el caso anterior, también fue considerado el punto de la oración anterior: punto + preposición + SN = SN SPINIC.
- Fuera argumento de SP, como por ejemplo: “en estudios epidemiológicos”.

Para los dos experimentos, fue realizada la limpieza de los términos extraídos conforme a lo explicado en la Sección 4. Fueron quitadas las stopwords de las extremidades de los candidatos y los candidatos que correspondían integralmente a stopwords. En este trabajo se tomó como base la lista de stopwords disponible en el Proyecto Snowball [29]. A esta lista se añadieron las conjugaciones de los verbos ‘poder’ y ‘deber’ y también algunas palabras como ‘año’, ‘día’, ‘misma’, etcétera, totalizando 733 palabras vacías.

Además, en este proceso fueron quitados también los numerales; los candidatos que son compuestos por apenas una letra cualquiera, etcétera. Así por ejemplo, un candidato a término como “los cuatro médicos de familia”, después de la limpieza, queda: “médico de familia”.

Asimismo, para los SN ADJUN se quitaron los svn de la extremidad de la derecha del candidato a término. Por ejemplo, para el SN ADJUN "se demostró la hiperreactividad bronquial", después de la remoción del svn 'se demostró' y de la limpieza, el candidato queda: "hiperreactividad bronquial".

Luego de la limpieza, los candidatos a términos fueron separados en listas de unigramas, bigramas, trigramas y mayores que trigramas para permitir su evaluación.

5.2. Ejecución y evaluación de los experimentos

Se realizaron dos evaluaciones automáticas distintas de los candidatos a términos extraídos, de acuerdo con la Figura 2. La primera, que utilizó la lista de términos de referencia del IULA, tuvo como objetivo verificar la calidad de los candidatos a términos extraídos. La segunda, que utilizó SNOMED CT ®, consistió en verificar si los candidatos que no estaban en la lista de términos de referencia, pero que fueron identificados manualmente como candidatos interesantes, podrían ser añadidos como términos del dominio de medicina.

En la primera evaluación se compararon los dos experimentos con sintagmas por medio de las métricas de precisión y cobertura. Debido a las variaciones morfológicas de las palabras (como plural y singular, masculino y femenino, etcétera), fue necesario hacer una simplificación de los términos extraídos. Lo mismo se hizo con las listas de términos de referencia, para que se las pudieran comparar ambas.

Para la simplificación, se eligió la técnica más radical de reducción de las palabras y la más utilizada en la literatura, que es el stemming. El stemming reduce las palabras a sus formas inflexionadas y no derivativas, es decir, elimina todo tipo de prefijos y sufijos, o "transforma" un verbo en su infinitivo [30], siendo cada palabra analizada aisladamente (por ejemplo: "protocolo de tratamiento" se simplifica en "protocol de tratamient").

Aquí se utilizó la herramienta PreTexT II [31], que aplica el stemming de acuerdo con el algoritmo de Porter utilizando el paradigma de orientación a los objetos en Perl. Esa herramienta puede ser aplicada el portugués, el inglés y el español, a la vez que tiene opciones estadísticas de selección de candidatos a términos.

Para la evaluación se utilizaron las métricas de precisión y cobertura.

En la Figura 3 se enseñan la precisión y la cobertura obtenidas a partir de los dos experimentos (SN, S SN). Esas se comparan con los resultados del trabajo que también ha utilizado el mismo corpus [15].

En la Figura 3, se presenta una comparación con los resultados de trabajo de Vivaldi y Rodríguez, en la cual EWN corresponde a la extracción de términos utilizando el método de YATE [24]. Los otros ítems corresponden a la extracción de términos utilizando las categorías de la Wikipedia (WP) con el nombre de dominio "Medicina" y variando el cálculo del coeficiente de dominio. El WP.lc es cuando se consideró el número de pasos simple recorridos en la Wikipedia; WP.lmc, cuando se consideró la media de la cantidad de los caminos recorridos en la Wikipedia; WP.nc, cuando se consideró el número de caminos recorridos en la Wikipedia. Es importante resaltar que la extracción de términos hecha por Vivaldi y Rodríguez tomó en cuenta solamente los candidatos que tenían el patrón de Sustantivo para los unigramas; de Sustantivo+Adjetivo para los bigramas; y de Sustantivo+Preposición+Sustantivo para los trigramas. Mientras la extracción de términos realizada en ese artículo consideró otras combinaciones posibles de los SN.

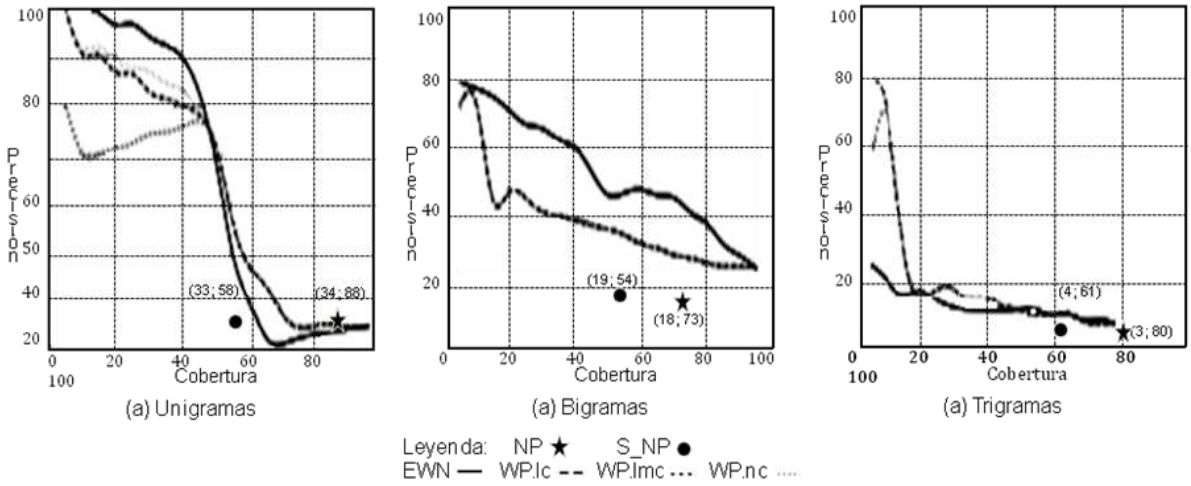

Figura 3. Precisiones y coberturas obtenidas. Modificación de [15]

Si comparamos los dos experimentos (SN y S SN) entre sí, se observa poca variación de las precisiones para unigramas, bigramas y trigramas. Por otro lado, las coberturas varían. Las mejores coberturas, obviamente, fueron obtenidas cuando se utilizaron todos los sintagmas nominales (SN).

En cambio, si comparamos a ambos con los resultados de Vivaldi y Rodríguez [15], para unigramas, los resultados de la extracción utilizando la subclasificación de SN establecida (Exp. S SN) fueron parecidos; para bigramas y trigramas, los resultados de Vivaldi y Rodríguez son superiores. No obstante, se deben hacer dos observaciones. La primera es que los resultados tienden a ser menores debido a que los experimentos de ese artículo no consideraron únicamente los patrones existentes en las listas de términos de referencia, sino que por el contrario, consideraron todas las posibilidades. La segunda observación es que los experimentos tienen una simplicidad para extraer términos porque trabajan solamente con los sintagmas nominales, puesto que no se necesita de conocimiento externo además del diccionario de la lengua que SMORPH utiliza y de las reglas de reagrupamiento de MPS.

Si bien la precisión no fue tan satisfactoria, fueron identificados manualmente candidatos a términos que no estaban en la lista de referencia, pero que parecían candidatos interesantes para el dominio. Vale aclarar además, que de la subclasificación de SN, los SN que formaban parte de una enumeración fueron los que mejor precisión arrojaron, ya que la mayoría se correspondían con los términos de las listas de referencia.

En la segunda evaluación se verificó la calidad de los candidatos en relación al dominio de medicina. Para a dicha tarea, se utilizó la lista de términos de SNOMED CT®, que contiene 1.060.632 términos en español. Al los términos de SNOMED CT® fue aplicado la misma limpieza de los candidatos a términos, o sea, fueron stemmizados, removidos acentos y cambiadas las palabras a sus formas minúsculas.

Luego, los candidatos a términos identificados manualmente se verificaron automáticamente, para saber si estaban presentes en la lista de términos del SNOMED CT®. Esa verificación fue realizada separadamente para cada experimento (Exp. SN y Exp. S SN) y los resultados están separados por (a) unigramas, (b) bigramas y (c) trigramas, conforme la Figura 4. En esa figura son mostrados los candidatos identificados que pueden ser considerados como términos de acuerdo con SNOMED CT®.

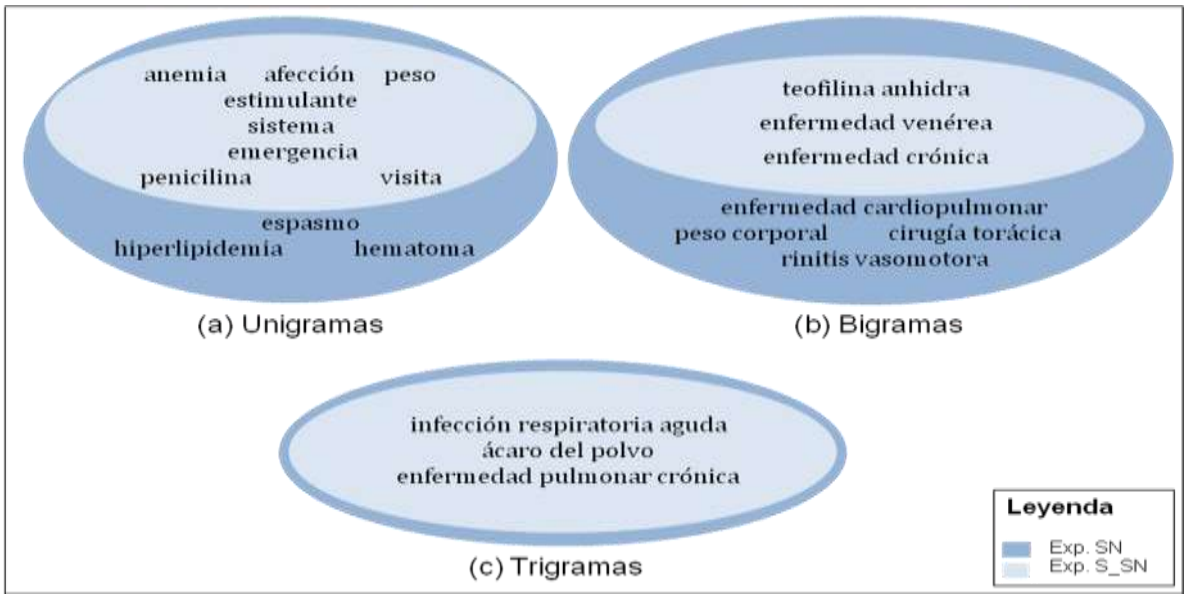

Figura 4. Términos extras obtenidos

Esa verificación posibilitó comprobar que esos candidatos son términos descriptivos del dominio de la medicina. Con eso, es posible colaborar con la lista de términos de referencia del IULA, sugiriéndoles la incorporación de dichos términos. Esa lista de referencia es importante para el área de medicina y también para los experimentos que pueden ser realizados con textos del dominio, porque permitiría una mejor evaluación de los resultados. Además, se observa la existencia de trigramas que no están en el patrón Sustantivo+Preposición+Sustantivo de la lista de referencia, como por ejemplo "infecciones respiratorias agudas" y "anticuerpos antimúsculo liso", lo que indica que hay trigramas importantes para el dominio que no necesariamente se ajusten a dicho patrón.

Además, de acuerdo con Krauthammer y Nenadic [18], se introducen nuevos términos en el vocabulario de medicina continuamente, por eso es imposible tener terminologías actualizadas de inmediato que sean producidas y supervisadas a mano.

Sobre la base de esa afirmación, se realizó una identificación manual de candidatos interesantes para el dominio y fueron observados trigramas que no están en la lista de referencia del corpus ni en SNOMED CT®, pero aún así, eran interesantes para el dominio de la medicina. En el Exp. SN hay ejemplos de esos casos: "*reacciones colaterales mínimas*", "*insuficiencia ventilatoria obstructiva*", "*patología respiratoria previa*", "*pacientes asmáticos atópicos*", "*factores potencialmente asmógenos*", "*enfermedades atópicas respiratorias*", "*diabetes insípida vasopresinsensible*", "*enfermedades cardiorrespiratorias crónicas*", "*insuficiencia ventilatoria súbita*", "*síndromes neurológicos infecciosos*", "*poliposis nasal bilateral*", "*músculo liso traqueobronquial*", "*intolerancia al medicamento*", "*rinitis alérgica estacional*" y "*traumatismo encefalocraneano*".

6. Consideraciones finales

La extracción de términos es una de las tareas más importantes y, a la vez, más complejas, en múltiples áreas y actividades de dominios específicos, como ser la construcción de diccionarios electrónicos, clasificación textual, resumen automático, creación de ontologías, etcétera. Esas actividades, indirectamente, ayudan a los especialistas de los dominios trabajados. A tales efectos, el objetivo del presente trabajo consistió en la comparación de la extracción de términos hecha de tres maneras diferentes. En la primera, se extrajeron todos los SN que estaban en el corpus; en las otras dos extracciones se consideraron diferentes SN destacados, que son los SN que corresponden

a ciertos criterios sintácticos, posicionales, etcétera. Para realizar tales extracciones, se utilizó el corpus IULA de medicina en español. Los resultados de las dos experimentaciones fueron comparados entre sí, a la vez que fueron evaluados automáticamente con una lista de términos de referencia y comparados con el trabajo de Vivaldi y Rodríguez [15], que utilizó el mismo corpus.

De acuerdo con los resultados arrojados por los dos experimentos, se pudo observar que los valores de cobertura variaron, mientras que los de precisión prácticamente se mostraron iguales. Como primera contribución de este artículo, puede decirse que, para el dominio de medicina, en español, es posible obtener mejores resultados de cobertura cuando se utilizan todos los SN, y van decreciendo a medida que la extracción se limita a SN específicos. Vale aclarar que se esperaba esta clase de resultados puesto que cuando se extraen todos los SN se obtiene una mayor cantidad de candidatos, lo que posibilita una mayor cobertura. No obstante, se esperaba que las precisiones de los experimentos con los SN determinados fueran mejores debido a que hubo una especificación en los patrones buscados como candidatos.

En relación con los resultados obtenidos por Vivaldi y Rodríguez, los nuestros fueron más bajos, solamente los resultados para los unigramas fueron similares. Una de las razones podría radicar en que los autores mencionados consideraron solo algunos patrones específicos, mientras que nosotros tomamos a todas las combinaciones posibles de sintagmas. A todo esto, es interesante destacar, sin embargo, la simplicidad utilizada para extraer los términos, ya que se trabajó únicamente con los SN, y para ello no se necesitan conocimientos externos aparte del diccionario que utiliza SMORPH y las reglas de reagrupamiento de MPS. Asimismo, se espera lograr mejoras en la precisión mediante el refinamiento de reglas de reagrupamiento de MPS, incluyendo otros elementos marginales que rodean al SN, como ser otros signos de puntuación, otros sintagmas, etcétera.

Por otro lado, y a modo de segunda contribución, se observaron términos interesantes que no estaban en la lista de referencia. Estos fueron evaluados en sus calidades con el SNOMED CT® y se concluyó que existen términos que podrían ser añadidos en la lista de referencia del IULA, lo que implica una mejora en esta. Asimismo, se espera que el perfeccionamiento de los recursos del dominio de medicina en español incentive las investigaciones en el área.

Por último, la tercera contribución radica en la ampliación del diccionario de SMORPH, en la medida en que se adicionaron 2.445 nuevos términos, en su mayoría, referidos al área de medicina. De esta forma, se espera que se mejoren los experimentos realizados en este dominio con dicho analizador. Como trabajo a futuro, se pretende mejorar la precisión de la extracción de términos a partir de nuevas reglas de filtrado de SN; continuar incrementando el diccionario de SMORPH, y probar las reglas de extracción automática en corpus más extensos, como así también en otros dominios.

Agradecimientos

Los autores agradecen a Erasmus Mundus, FAPESP, CNPq y CONICET por el apoyo financiero y a Vivaldi y Rodríguez por facilitarnos el corpus y las listas de referencias utilizados.

Referencias

[0] Este trabajo es una versión extendida del presentado en RANLP'2011 Recent Advances in Natural Language Processing, Bulgaria, con el título "Experiments on Term Extraction using Noun Phrase Subclassifications". http://lml.bas.bg/ranlp2011/proceedings.php

[1] Corpus Técnico IULA-UPF. "Datos procedentes del CORPUS TÉCNICO del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra

(http://bwananet.iula.upf.edu/) en el período (diciembre/2010)".

[2] Aït-Mokhtar, S. L'analyse présintaxique en une seule étape. Tesis doctoral. Universidad BlaisePascal/Grilll, Clermont-Ferrand, 1998.

[3] Abbaci, F. Développment du Module Post-Smorph. Memória del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL. Clermont-Fd.

[4] SNOMED CT®: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. "This material includes SNOMED Clinical Terms R (SNOMED CT®) which is used by permission of the International Health Terminology Standards Development Organisation (IHTSDO). All rights reserved. SNOMED CT®, was originally created by The College of American Pathologists. 'SNOMED' and 'SNOMED CT®' are registered trademarks of the IHTSDO".

[5] Sager, J. Curso práctico sobre el procesamiento de la terminología. Madrid: Fundación Germán Sánchez Ruizpérez, 1993.

[6] Barrón-Cedeño A.; Sierra G.; Drouin P.; Ananiadou S. An improved automatic term recognition method for Spanish. In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 5449 of Lecture Notes in Computer Science, pages 125-136. Springer Berlin / Heidelberg, 2009.

[7] Bosma, W.; Vossen. P. Bootstrapping language neutral term extraction. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.

[8] Bonin, F.; Dell'Orletta F.; Montemagni S.; Venturi G. A contrastive approach to multi-word extraction from domain-specific corpora. In K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.

[9] A. Gelbukh, G.; Sidorov, E. Lavin-Villa, L. Chanona-Hernandez. Automatic term extraction using log-likelihood based comparison with general reference corpus. In Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, NLDB'10, pages 248-255, Berlin, Heidelberg, 2010. Springer-Verlag.

[10] Castro, E. Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts. In Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, pages 751-757, New York, NY, USA, 2010. ACM.

[11] Lacoste, C.; Joo-Hwee Lim; Chevallet, J.-P.; Le, D.T.H. Medical-image retrieval based on knowledge-assisted text and image indexing. IEEE Circuits and Systems for Video Technology 17(7):889-900, 2007.

[12] Sánnchez, D; Batet M.; Valls A. Web-based semantic similarity: An evaluation in the biomedical domain. Int. J. Software and Informatics, 4(1):39-52, 2010.

[13] López Rodríguez, 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, 2(8), 2006.

[14] López Rodríguez, C.; Benítez P.; Sánchez M. Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm. Panace, VII (24):228-240. 2006.

[15] Vivaldi, J. and Rodríguez, H. Using wikipedia for term extraction in the biomedical domain: First experiences. Procesamiento del Lenguaje Natural, 45:251-254, 2010.

[16] Vivaldi, J.; Cunha I.; Torres-Moreno J.; Velazquez-Morales P. Automatic summarization using terminological and semantic resources. K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.

[17] Alarcón, R. Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios. In [CD-ROM], editor, Serie Tesis, number 26. Barcelona: IULA, 2010. ISBN: 13: 978-84-89782-46-4.

[18] Krauthammer, M.; Nenadic, G. Term identification in the biomedical literature. J. of Biomedical Informatics, 37:512-526, December 2004.

[19] Martín Mingorance. Modelo Lexemático Funcional. Universidad de Granada, 1 edition, 1998.

[20] Martínez, S. Estructuración conceptual y formalización terminográfica de frasemas en el subdominio de la oncología. Estudios de Lingüística del Español, 19(19), 2003.

[21] Sierra, G.; Alarcon, R.; Molina, A.; Aldana, E. Web exploitation for definition extraction. In Proceedings of the 2009 Latin American Web Congress, LA-WEB'09, pages 217-223, Washington, DC, USA, 2009. IEEE Computer Society.

[22] Alarcón, R. Extracción automática de contextos definitorios en corpus especializados. PhD thesis, Universidad Pompeu Fabra, Barcelona, 2009.

[23] Sierra, G. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. Linguamática, 1(2):13-37, December 2009.

[24] Vivaldi, J. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD thesis, Universitat Politécnica de Catalunya, Barcelona, España, 2001.

[25] EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

[26] Moreno-Sandoval, A. Terminología y sociedad del conocimiento. 2009.

[27] Abney, S. Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny, editors, Principle-Based Parsing, chapter 1, pages 1-18. Kluwer Academic Publishers, Dordrecht, 1991.

[28] Referencias: EMS: etiqueta Morfosintáctica; nom: nombre; GEN: género; fem: femenino; NUM: número; PL: plural; v: verbo; ind: indicativo; PERS: persona; 2a: segunda, TPO: tiempo; pres: presente; TR: tipo de regularidad; irr: irregular; TC: tipo de conjugación; c1: primera conjugación; TDIAL: tipo de variedad dialectal; est: estándar; prde: preposición de; prep: preposición; masc: masculino; cop: conjunción copulativa; sg: singular, linsing: línea siguiente; pun: punto.

[29] Lista de stopwords utilizada de: http://snowball.tartarus.org/algorithms/spanish/stop.txt

[30] Manning, C.; Raghavan P., Schütze H.. Language models for information retrieval. In An Introduction to Information Retrieval, chapter 12. Cambridge University Press, 2008.

[31] Soares, M.; Prati, R.; Monard, M. Pretext II: Descrição da reestructuração da ferramenta de pré-processamento de textos. Technical Report 333, Instituto de Ciencias Matemáticase de Computação (ICMC) - USP - São Carlos, São Carlos - SP, 2008.