Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus

Recognition of Terms in Spanish by Applying a Contrastive Approach

Olga Acosta
Departamento de Ciencias del Lenguaje
Pontificia Universidad Católica de Chile
oacostal@uc.cl

César Aguilar
Departamento de Ciencias del Lenguaje
Pontificia Universidad Católica de Chile
caguilara@uc.cl

Tomás Infante
Magíster en Procesamiento y Gestión de la Información
Pontificia Universidad Católica de Chile
tomasinfante@gmail.com

Resumen

En este artículo presentamos una metodología para la identificación y extracción de términos a partir de fuentes textuales en español correspondientes a dominios de conocimiento especializados mediante un enfoque de contraste entre corpus. El enfoque de contraste entre corpus hace uso de medidas para asignar relevancia a palabras que ocurren tanto en el corpus de dominio como en corpus de lengua general o de otro dominio diferente al de interés. Dado lo anterior, en este trabajo realizamos una exploración de cuatro medidas usadas para asignar relevancia a palabras con el objetivo de incorporar la de mejor desempeño a nuestra metodología. Los resultados obtenidos muestran un desempeño mejor de las medidas diferencia de rangos y razón de frecuencias relativas comparado con la razón *log-likelihood* y la medida usada en Termostat.

Palabras clave

Término, *unithood*, *termhood*, extracción terminológica, lenguaje especial

Abstract

In this article we present a methodology for identifying and extracting terms from text sources in Spanish corresponding specialized-domain corpus by means of a contrastive approach. The contrastive approach requires a measure for assigning relevance to words occurring both in domain corpus and reference corpus. Therefore, in this work we explored four measures used for assigning relevance to words with the goal of incorporating the best measure in our methodology. Our results show a better performance of rank difference and relative frequency ratio measures compared with *log-likelihood* ratio and the measure used by Termostat.

Keywords

Term, *unithood*, *termhood*, term extraction, special language

1 Introducción

Desde el punto de vista del aprendizaje de ontologías a partir de textos, el reconocimiento automático de términos (en inglés, ATR) se considera un prerrequisito para tareas más complejas como son, por ejemplo, la extracción de conceptos y taxonomías (Buitelaar et al., 2005). A grandes rasgos, un término es una representación lingüística de conceptos de dominio específico (Kageura & Umino, 1996; Pazienza, 1998; Vivaldi et al., 2001), y una terminología construida de forma coherente puede, por tanto, ser útil como plataforma básica para construir ontologías y además usada para otras aplicaciones importantes (diccionarios, traducción automática, indexación, tesauros, etc.). En este sentido Pazienza et al. (2005) señalan también el valor que tiene la extracción automática de términos como punto de partida para desarrollar sistemas inteligentes y con ello mitigar el cuello de botella en la adquisición de conocimiento. Los enfoques usados para la extracción automática de términos son los siguientes: i) lingüístico, ii) estadístico, iii) aprendizaje automático, y iv) métodos híbridos (Ananiadou & Mcnaught, 2005; Lossio-Ventura et al., 2014; Kockaert & Steurs, 2015). Por un lado, se han definido técnicas lingüísticas para filtrar candidatos no relevantes, por ejemplo, vía la configuración de patrones morfosintácticos, mientras que la parte estadística y/o pro-

babilística conlleva la aplicación de medidas estadísticas para asignar relevancia a términos candidatos. Por otro lado, los enfoques de aprendizaje automático (en inglés, ML) usan datos de entrenamiento para aprender rasgos útiles para la extracción de términos. Finalmente, los métodos actuales consisten en híbridos que incorporan algunos o potencialmente todos los enfoques anteriores para identificar y reconocer términos (Vivaldi & Rodríguez, 2007). Los enfoques actuales están basados preponderantemente en métodos probabilísticos y lingüísticos debido a que el principal reto en aprendizaje automático es seleccionar un conjunto de rasgos discriminantes que caractericen los términos (Lossio-Ventura et al., 2014), lo que representa una tarea compleja.

La enorme cantidad de información digital disponible y áreas de conocimiento que evolucionan rápidamente, como es el caso de la biomedicina (Kageura & Umino, 1996; Poesio, 2005; Ananiadou & Mcnaught, 2005), influyen directamente en el interés por mejorar los métodos actuales para la extracción automática de términos e implementarlos en sistemas computacionales con la meta de agilizar el trabajo de identificación y extracción del vocabulario de un dominio. Los rasgos o propiedades que caracterizan a los términos son *unithood* y *termhood*, tal como lo proponen Kageura & Umino (1996). Estos rasgos se han explorado en la literatura sobre la extracción automática de términos (Vivaldi et al., 2001; Ananiadou & Mcnaught, 2005; Kit & Liu, 2008; Barrón-Cedeño et al., 2009; Gelbukh et al., 2010; Spasić et al., 2013; Kockaert & Steurs, 2015). Asimismo, se han considerado otras cuestiones relevantes, como es el caso de la ambigüedad y variación de los términos (Daille et al., 1996; Spasić et al., 2013), que dependiendo de la aplicación para la que se realiza la extracción de terminología adquirirán mayor o menor importancia (Kockaert & Steurs, 2015). Con relación a las variantes de términos, tanto Ananiadou (1994) como Vivaldi & Rodríguez (2007) señalan que a los términos se les atribuyen rasgos de no ambigüedad y mono-referencialidad para designar conceptos en un dominio, sin embargo, esto dista mucho de la realidad debido a que los problemas de polisemia, homonimia y sinonimia ocurren con más frecuencia de lo esperado.

En este artículo presentamos una metodología para reconocer y extraer términos candidatos, tanto unipalabra como multipalabra. Nuestra propuesta consiste en un enfoque de comparación entre corpus para identificar las palabras relevantes del dominio analizado, así como asignarles una ponderación que refleje su relevancia. Para lograr lo anterior, comparamos cuatro medidas diferentes para calcular el *termhood* de cada palabra común al corpus de dominio y de referencia: razón de frecuencia relativa (Manning & Schütze, 1999), razón log-likelihood (Gelbukh et al., 2010), diferencia de rangos (Kit & Liu, 2008) y la medida usada en *TermoStat* (Drouin, 2003). Bajo este enfoque contrastivo de corpus, asumimos que una palabra estrechamente relacionada con el dominio debe tener una probabilidad de ocurrencia más alta en dicho dominio que en un corpus de referencia. Así, si este proceso de asignación de relevancia es eficaz, palabras de dominio tendrán ponderaciones mayores que palabras no relacionadas con el dominio. En una fase posterior, la relevancia de la palabra se puede usar para extraer términos candidatos multipalabra, de modo que las palabras con ponderaciones altas contribuirán a incrementar la relevancia de sintagmas nominales cuando están presentes (*termhood* multipalabra). En el caso de la propiedad de *unithood*, consideramos que los patrones morfosintácticos constituyen una buena evidencia de *unithood* (Vivaldi & Rodríguez, 2007; Kockaert & Steurs, 2015). Además, como parte de la metodología, proponemos también la implementación de heurísticas lingüísticas para construir automáticamente una lista de adjetivos no relevantes del dominio analizado. Esto último es relevante ya que adjetivos (principalmente adjetivos relacionales) tienen una interpretación composicional, por lo que medidas tradicionales (por ejemplo, información mutua) fallan en la tarea de mostrar *unithood* de términos candidatos multipalabra.

En la sección 2 presentamos trabajos relacionados con la extracción automática de términos. En la sección 3, discutimos algunas cuestiones relacionadas con términos y su comportamiento. En la sección 4, describimos nuestra propuesta metodológica para la extracción automática de términos. En la sección 5 presentamos resultados de nuestros experimentos. Finalmente, en la sección 7 bosquejamos nuestras conclusiones.

2 Trabajo relacionado

La extracción automática de términos se ha utilizado para construir recursos lexicográficos como diccionarios, glosarios y vocabularios, así como recursos computacionales que sean útiles en el procesamiento automático de textos. Asimismo, tareas como la recuperación de información, clasificación textual, traducción automática, etc., se han beneficiado de los avances en la extracción automática de términos. Fi-

nalmente, si asumimos que los términos denotan conceptos, una terminología bien construida puede ser un punto de partida importante para el aprendizaje de ontologías (Kageura & Umino, 1996; Buitelaar et al., 2005; Poesio, 2005; Wong, 2008; Kockaert & Steurs, 2015).

Como se mencionó en párrafos anteriores, existen por lo menos tres enfoques diferentes para la extracción automática de terminología, sin embargo, ninguno considerado de forma independiente, ha sido completamente exitoso. Por un lado, los enfoques lingüísticos o basados en reglas intentan filtrar términos candidatos mediante patrones de formación de términos como evidencia de *unithood* (Ananiadou, 1994; Justeson & Katz, 1995; Bourigault et al., 1996; Heid, 1998; Jacquemin, 2001). Debido a que el uso de patrones morfosintácticos no ayuda a discernir entre palabras de dominio y de uso general, el enfoque común es generar una lista de palabras vacías (en inglés, *stopword list*) como una forma de filtrar candidatos no relacionados con el dominio. Una desventaja importante del enfoque lingüístico es la gran cantidad de ruido que produce y el hecho de que no es directamente aplicable a diferentes dominios y lenguajes.

Los enfoques estadísticos, por otro lado, usan un número diferente de medidas y distribuciones estadísticas para calcular *unithood* y *termhood*. En el caso específico de *unithood*, este tipo de enfoques considera dos propiedades que son comunes en términos multipalabra: combinaciones de palabras relativamente estables y cuya ocurrencia es alta. Los enfoques estadísticos, dada esta estabilidad sintagmática, y la variación nula en el orden de las palabras, pueden enfocarse en analizar n-gramas sin considerar la estructura lingüística subyacente. Las medidas usadas para el cálculo de colocaciones representan un buen ejemplo de cálculo de *unithood*. En términos formales, estas medidas cuantifican cuánto se desvía lo observado de lo que se espera como producto del azar, dadas las frecuencias individuales de las palabras. Entre las medidas para cuantificar la divergencia entre lo observado y esperado están el estadístico X^2 usado en Drouin (2003), así como en Matsuo & Ishizuka (2004). Otras medidas para el cálculo de *unithood* incluyen el puntaje-t (*t-score*), razón *log-likelihood* (Dunning, 1993), información mutua (Church & Hanks, 1990) y el coeficiente phi. Por su parte, Wermter & Hahn (2005) consideran el grado de remplazamiento de las palabras constituyentes de un término multipalabra por otras (modificabilidad paradigmática). Existen pocos ejemplos de enfoques donde únicamente se aplique la estadística para el proceso de extracción automática de términos, generalmente las medidas para el cálculo de *unithood* se combinan con una fase lingüística y se calculan para las combinaciones que pasaron ya el filtro lingüístico. Un experimento que sólo considera una fase estadística es el realizado por Pantel & Lin (2001).

Respecto a la medición del rasgo *termhood*, que refiere al grado en que una unidad léxica denota conceptos del dominio, el enfoque inicial fue el uso de la frecuencia del término candidato en el dominio como un indicio de su importancia (Daille et al., 1994). Algunos autores fijan el origen de la extracción automática de términos al campo de la recuperación de información y esta relación estrecha se evidencia al considerar medidas como el TF-IDF para el cálculo de *termhood* (Evans & Lefferts, 1995; Medelyan & Witten, 2006). En este sentido, se aplica la fórmula TF-IDF para ponderar alto aquellos candidatos con un nivel de especificidad mayor. Un tercer método se enfoca en el uso contextual de los candidatos a término que ya pasaron un filtro lingüístico para después analizar su co-ocurrencia con palabras de contexto adicionales (Maynard & Ananiadou, 1999; Frantzi et al., 2000). Un cuarto método se enfoca en los candidatos unipalabra y analiza su estructura morfológica interna (Aubin & Hamon, 2006). Áreas de conocimiento, como la medicina, derivan en gran parte su terminología de raíces griegas y latinas, lo que se puede explotar como un rasgo de *termhood* (Ananiadou, 1994). Finalmente, otro enfoque consiste en contrastar el comportamiento de un candidato dentro del dominio con información de un corpus de referencia o lengua general. En este método se asume que los términos son específicos de dominio, y como consecuencia ocurren con mayor frecuencia en su dominio que en otros dominios o en lengua general. Bajo esta premisa, se compara la frecuencia de un candidato en un corpus de dominio específico con su frecuencia en un corpus de referencia o de otro dominio diferente. Por ejemplo, el método *contrastive weight* de Basili et al. (1997) es una adaptación del TF-IDF porque en lugar de considerar la ocurrencia en diferentes documentos de una colección se usa la dispersión de los candidatos en dominios diferentes. Por su parte Ahmad et al. (1999) usan una medida para hacer referencia al concepto de *weirdness* de una palabra mediante la comparación de las frecuencias normalizadas de la palabra entre un corpus especializado y uno de referencia. Chung (2003) usa una razón de frecuencia normalizada para medir *termhood*. Por su parte, Wong (2008) usa el comportamiento distribucional de una palabra en otro corpus para medir la distribución

intra-dominio y el comportamiento distribucional multi-dominio. Drouin (2003) compara precisión y cobertura para la clasificación de métodos de prueba de hipótesis diferentes, en un intento por determinar el mejor método. Por último, Kit & Liu (2008) y Gelbukh et al. (2010) miden *termhood* de palabras simples mediante la medida de diferencia de rangos y la razón *log-likelihood*, respectivamente.

Finalmente, los sistemas de aprendizaje automático usan datos de entrenamiento para aprender rasgos que sean útiles y relevantes para el reconocimiento y clasificación de términos. Varias técnicas de aprendizaje automático se han usado para identificar y clasificar términos, los que incluyen HMMs (Collier et al., 2000), enfoques Bayesianos, SVMs (Kazama et al., 2002; Yamamoto et al., 2003), y árboles de decisión (Lopez & Romary, 2010).

3 Términos y su comportamiento

Los términos son palabras o unidades léxicas que denotan conceptos en un dominio restringido (Ananiadou, 1994; Daille, 1996; Jacquemin, 1997; Pazienza, 1998; Frantzi et al., 2000; Vivaldi et al., 2001; Buitelaar et al., 2005; Wong, 2008; Spasić et al., 2013). De acuerdo con Kageura & Umino (1996) y Daille et al. (1996), los términos son expresiones principalmente multipalabra de tipo nominal, caracterizadas por propiedades morfológicas, sintácticas y semánticas. Términos como *dominio restringido, lenguaje de especialidad, lenguaje especial, sublenguaje, dominio especializado, dominio especialista*, por otro lado, refieren a un subsistema lingüístico con términos especializados y otros recursos lingüísticos usados para comunicar de manera precisa y sin ambigüedad en un área de conocimiento determinada (Vivanco, 2006; Ananiadou, 1994).

Aunque los términos son representaciones lingüísticas para denotar conceptos en dominios especializados, no es posible distinguirlos completamente de palabras comunes por su forma debido a que los lenguajes de especialidad se derivan del lenguaje general, y por ello siguen las mismas reglas de formación de palabras (L'Homme, 2004). No obstante, Pazienza et al. (2005), señalan que hay un gran interés dentro de la lingüística computacional por establecer una definición más profunda de lo que es un término, con el fin de desarrollar algoritmos para mejorar el desempeño de los enfoques actuales de extracción.

Con la meta de aclarar lo que es un término y aquello que se requiere para su identificación y extracción, se han propuesto dos propiedades: *unithood* y *termhood* (Kageura & Umino, 1996). La propiedad de *unithood* refiere al grado de estabilidad sintagmática de un candidato a término y es relevante solo para el caso términos multipalabra. Por otro lado, *termhood* refiere al grado de relevancia de un candidato al dominio y se enfoca en ambos, unipalabra y multipalabra. Particularmente, en la propiedad de *unithood*, estamos de acuerdo con Kit & Liu (2008) respecto a que es una condición necesaria pero no suficiente, ya que un término verdadero debe tener un *unithood* alto. Empero, pueden existir muchos candidatos que lo tengan, pero no por ello serán términos verdaderos.

Kageura & Umino (1996) mencionan que el origen de la extracción automática de términos se puede encontrar en el campo de la recuperación de información. En la recuperación de información, los términos índice o palabras clave se usan para indexar o recuperar documentos. Estos términos índice tienen algún significado en sí mismos, y regularmente tienen la categoría gramatical de nombres (Baeza & Rivera, 2011). En este escenario, no todos los nombres son relevantes para indexar documentos, es decir, algunos representan mejor los documentos que otros y pueden discriminar más efectivamente entre ellos. En el caso de la extracción automática de términos, no todos los nombres denotan conceptos relevantes en el dominio, por lo que es necesario asignar relevancia más alta a los más significativos que a los menos importantes, lo cual no es una tarea fácil.

3.1 Estructura de términos

De acuerdo con Daille (1996), Kageura & Umino (1996), así como Spasić et al. (2013), para propósitos prácticos, los términos se definen como sintagmas nominales que ocurren con frecuencia en textos de un dominio específico donde tienen un significado especial. En la extracción automática de términos para el español (Vivaldi, 2004; Vivaldi & Rodríguez, 2007), unidades como los nombres, adjetivos y la preposición *de*, son los más comunes que participan en la formación de términos multipalabra. Vivaldi & Rodríguez (2007) presentan datos respecto al uso de estos patrones en lenguajes de especialidad de dos subcorpus en español del corpus técnico del IULA. De 2,145 términos en estos dos subcorpus, 48 % son nombres únicos; 45 % son sintagmas nominales multipalabra, es decir, `<nombre+adjetivo>`, y 7 % tienen el patrón `<nombre+preposición+nombre>`, donde la preposición de tiene un uso mucho mayor que el resto

de las preposiciones. A partir de los datos se puede ver que, por lo menos en español, los términos de una sola palabra constituyen un grupo importante, contrario a la observación hecha por Daille et al. (1996) respecto a que la mayoría de los términos son sintagmas nominales multipalabra.

3.2 Patrones con frase preposicional

Daille et al. (1996) argumenta que los términos son secuencias que muestran diferentes tipos de variaciones, contrario a la concepción tradicional de que los términos son secuencias fijas. Para estos autores, una variante de un término es un enunciado, que es semántica y conceptualmente relacionado a un término original. Las variaciones se dividen en dos clases principales: morfológicas y sintácticas. Por ejemplo, una variante morfológica de *célula epitelial* es *célula asociada con el epitelio*. Por otro lado, una variación sintáctica de la misma expresión es *célula de tumor epitelial*. Adicionalmente, las variaciones sintácticas se pueden dividir en variaciones que preservan significado: célula sanguínea — *célula de la sangre*, y aquellas que incluyen un cambio en significado: *célula sanguínea — célula mononuclear sanguínea*. En el caso de variaciones que preservan significado, Daille et al. (1996) argumenta:

> The most constant signs of permutation occur around the preposition of. It is with this preposition, as opposed to the others, that a strict permutation without insertion leads to the best results, i.e., less noise. We can verify that the terms variance analysis or chromosome duplication are completely equivalent to textual sequences under which they have been identified (analysis of variance, duplication of chromosome). From this point of view, the sole permutation demonstrates a synonymous connection between the term in its basic form and its transformed form (p. 222).

En lo que respecta al uso de preposiciones, Marchis (2010) señala que los compuestos nominales (por ejemplo, *kidney diseases*) difícilmente se presentan en lenguas romance (por lo menos en rumano y español), por lo que se pueden combinar los nombres con preposiciones como *de* o *a*, sin embargo, los nombres con adjetivos relacionales se priorizan para evitar el abuso en el uso de frases preposicionales. Por otro lado, como Vivanco (2006) menciona, las preposiciones inherentes a las lenguas romances, que se usan con frecuencia en términos, aunque van en contra de la brevedad, sirven como un factor aclaratorio en español, esto podría, como Marchis (2010) sugiere, ser una causa importante de la alta productividad de frases nominales como `<nombre+adjetivo>`.

De acuerdo con Daille et al. (1996), las preposiciones usadas en las variantes de permutación se dividen en dos categorías: preposiciones *de, con, para*, y *en*, que sólo tienen una función relacional y el resto que son preposiciones locativas, de las cuales de es la preposición semánticamente menos informativa debido a que representa una gran cantidad de relaciones. En el cuadro 1, mostramos datos sobre el uso de preposiciones en frases nominales con estructura:

```
<Nombre + adjetivo? + preposición +
      determinante? + nombre>
```

en tres colecciones de textos diferentes: Corpus de Ingeniería Lingüística (Medina et al., 2004), textos extraídos de MedlinePlus en español y un corpus de referencia general recolectado automáticamente de la Web (extraído de un periódico mexicano). Como se puede observar en el cuadro 1, las preposiciones que tienen una función relacional representan más del 85 % de uso comparado con el resto. La preposición *de* en español también es por mucho la preposición más usada.

Preposición[1]	**Referencia** %	**CLI** %	**MedlinePlus** %
De	68.3	72.8	65.3
En	11.7	10.6	14
Con	3.4	3.3	6.5
Para	2.0	3.2	3.1
Resto	14.6	10.1	11.1

Cuadro 1: Uso de preposiciones en español.

3.3 Adjetivos no relevantes en terminología

Un adjetivo es una categoría gramatical cuya función es modificar nombres (Demonte, 1999; Fábregas, 2007). Existen dos tipos de adjetivos que asignan propiedades a los nombres: adjetivos calificativos o descriptivos y adjetivos relacionales. Los adjetivos calificativos refieren a rasgos constitutivos del nombre modificado. Estos rasgos se exhiben o caracterizan por medio de una propiedad física única: color, forma, predisposición..., (*el libro azul, la señora delgada*). Por otro lado, los adjetivos relacionales asignan un conjunto de propiedades, es decir, todas las características que conjuntamente definen nombres

[1] Las preposiciones consideradas por las Real Academia de la Lengua Española son: *a, ante, bajo, cabe, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre* y *tras*.

(*puerto marítimo, paseo campestre*). En terminología, los adjetivos relacionales representan un elemento importante para construir términos especializados, por ejemplo: *hernia inguinal, enfermedad venérea, desorden psicológico*, se consideran términos en medicina. En contraste, *hernia rara, enfermedad seria*, y *desorden crítico* parecen juicios más descriptivos y estrechamente relacionados con un contexto específico.

Identificación sintáctica de adjetivos no relevantes

Con base en lo anterior, si consideramos la estructura interna de adjetivos, se pueden identificar dos tipos: adjetivos permanentes y episódicos (Demonte, 1999). El primer tipo de adjetivo representa situaciones estables, propiedades permanentes que caracterizan individuales. Estos adjetivos se ubican fuera de cualquier restricción espacial o temporal (*psicópata, egocéntrico, apto*). Por otro lado, los adjetivos episódicos refieren a situaciones transitorias o propiedades que implican cambio y con limitaciones de espacio-tiempo. Casi todos los adjetivos descriptivos derivados de participios pertenecen a esta última clase, así como participios adjetivales (*harto, limpio, seco*). El español es uno de los pocos lenguajes que en su sintaxis representa esta diferencia en el significado de adjetivos. En muchos lenguajes esta diferencia solo es reconocible a través de la interpretación. En español, las propiedades individuales se predican con el verbo *ser*, y las episódicas con el verbo *estar*, lo que es esencial para probar a qué clase pertenece un adjetivo. Con la meta de identificar y extraer adjetivos no relevantes, proponemos extraer los adjetivos predicados con el verbo *estar*.

Otra heurística lingüística para identificar adjetivos descriptivos es que solo estos tipos de adjetivos aceptan adverbios de grado, y pueden ser parte de construcciones comparativas, por ejemplo, *muy alto, extremadamente grave*. Finalmente, solo los adjetivos calificativos pueden preceder un nombre porque —en español— los adjetivos relacionales siempre se posponen (*la antigua casa*).

4 Metodología

En este trabajo proponemos una metodología para extraer términos de un corpus de dominio especializado. La entrada debe ser un corpus con etiquetado morfosintáctico. En este caso, el corpus se ha etiquetado con FreeLing (Carreras et al., 2004). Los etiquetadores más usados para español son TreeTagger Schmid (1994) y FreeLing. En este experimento usamos FreeLing porque es más preciso. Lo cuadro 2 muestra las etiquetas más usadas.

Etiqueta	**Significado**
N.* (NC, NP)	Nombre
A.* (AQ, AO)	Adjetivo
V.* (VM, VA, VS)	Verbo
RG	Adverbio
SP	Preposición
D.* (DD, DP,DA,DI, DT,DE)	Determinante
P.* (PP, PD, PX, PI, PT, PR, PE)	Pronombre
C.* (CC, CS)	Conjunción
F.* (FA, FC, FZ, FG, FS)	Puntuación

Cuadro 2: Las etiquetas FreeLing más comunes.

4.1 Estandarización de etiquetas

La mayoría de las etiquetas fueron truncadas a dos caracteres, excepto en el caso del verbo *estar*, cuya etiqueta es VAE.

4.2 Análisis sintáctico superficial

El análisis sintáctico superficial es el proceso de identificar y clasificar segmentos de una oración vía la agrupación de las etiquetas morfosintácticas principales que forman frases no recursivas básicas. Una gramática para el análisis sintáctico superficial es un conjunto de reglas que indican cómo deben agruparse las oraciones. Las reglas de una gramática usan los patrones de etiquetas para describir secuencias de palabras etiquetadas, por ejemplo, `<DA>?<NC><AQ>*`. Los patrones de etiquetas son similares a patrones de expresión regular, donde los símbolos como "`*`" significan cero o más ocurrencias, "`+`" significa una o más ocurrencias y "`?`" representa un elemento opcional.

En este trabajo nos enfocamos en la extracción de términos base a partir de información textual. Lingüísticamente, como sucede en inglés, los patrones de términos más productivos consisten de un nombre y cero o más adjetivos (Vivaldi et al., 2001; Barrón-Cedeño et al., 2009). Estos términos se denominan términos base, de los cuales se derivan términos más complejos (Daille et al., 1996). Mediante el uso de etiquetas FreeLing, estos patrones se pueden representar como una expresión regular:

`<NC><AQ>*`

Como Daille et al. (1996) mencionan, si consideramos el patrón:

`<nombre><preposición De><nombre>`

En muchos casos, los términos con una frase preposicional con el núcleo *de* son variantes de formas básicas, en este caso, del patrón `<nombre><adjetivo>` (por ejemplo, *enfermedad renal-enfermedad del riñón*). Finalmente, la expresión regular usada para extraer adjetivos no relevantes de acuerdo con las heurísticas mencionadas en la sección 3 son:

```
<RG><AQ>
<VAE><AQ>
<D.*|P.*|F.* |S.*><AQ><NOUN>
```

Donde RG, AQ y VAE, corresponden a las etiquetas para adverbios, adjetivos, y el verbo estar, respectivamente. Las etiquetas `<D.*|P.*|F.*|S.*>` corresponden a determinantes, pronombres, signos de puntuación y preposiciones. La expresión `<D.*|P.*|F.*|S.*>` es una restricción para reducir *ruido*, ya que elementos erróneamente etiquetados por FreeLing como adjetivos se extraen sin esta restricción.

4.3 Reducción de ruido

Con el objetivo de eliminar palabras no relevantes de frases nominales antes de asignar relevancia a términos candidatos multipalabra utilizamos los adjetivos descriptivos obtenidos mediante heurísticas lingüísticas. Sumado a lo anterior, los sintagmas candidatos que tienen como núcleo nombres muy comunes como: *caso, mayoría, vez, superficie, área, tamaño, tipo, subtipo, forma, parte, término, clase* y *subclase*, son eliminados en esta fase.

Adjetivos no relevantes

De acuerdo con Pazienza et al. (2005), es posible utilizar un conjunto de palabras no relevantes al dominio para refinar la terminología que se deriva de un proceso automático. Barrón-Cedeño et al. (2009) considera una lista de palabras vacías con alta frecuencia en un corpus que se espera no formen parte de términos en un dominio específico. Consideramos que una lista construida de esta forma tiene desventajas debido a que además de la selección por frecuencia de ocurrencia se requiere la supervisión humana para determinar si una palabra es relevante o no al dominio.

Dado lo anterior, consideramos que la frecuencia de una palabra no basta como indicador y que se pueden tomar en cuenta heurísticas lingüísticas que operan en un lenguaje específico para automatizar la selección de palabras no relevantes dentro del dominio, sin embargo, una de las desventajas es que esto conduce a dependencia del lenguaje. En el caso del español, Demonte (1999) propone un conjunto de rasgos característicos para distinguir entre adjetivos calificativos y relacionales. Estas heurísticas se mencionaron en la sección 3.

Consideramos cómo se usan los adjetivos en el corpus de dominio en lugar de su frecuencia de uso, por lo que las heurísticas implementadas en Acosta et al. (2013) —un adverbio que precede a un adjetivo, un adjetivo que precede a un nombre, y el verbo *estar* precediendo un adjetivo— se usan para extraer adjetivos no relevantes del dominio como se mencionó en párrafos anteriores con relación a la distinción entre adjetivos permanentes y episódicos.

Finalmente, los adjetivos del corpus de referencia se obtuvieron también con estas tres heurísticas en mente y fueron manualmente revisados para determinar su relevancia a cualquier dominio de conocimiento especializado (por ejemplo, adjetivos como *relevante, importante, necesario, apropiado, correspondiente*, etc., se consideraron en la lista de adjetivos no relevantes). Esta es una lista de tamaño fijo que puede considerarse como lista base donde se pueden agregar los adjetivos extraídos del dominio. En el apéndice A presentamos un subconjunto de los mejores términos candidatos multipalabra, antes y después de reducir *ruido*. Como se puede observar de este subconjunto de candidatos antes de remover *ruido*, existen juicios descriptivos que están estrechamente relacionados con un contexto específico. Estos casos se recuperan con los patrones morfosintácticos implementados, por lo que es necesario aplicar una fase de reducción de ruido. Los adjetivos extraídos del dominio con heurísticas lingüísticas se presentan en el apéndice A.

4.4 Relevancia de palabras

Siguiendo el enfoque propuesto por Enguehard & Pantera (1995), primero evaluamos *termhood* de palabras simples con cuatro medidas propuestas en la literatura sobre la extracción automática de términos (Manning & Schütze, 1999; Drouin, 2003; Kit & Liu, 2008; Gelbukh et al., 2010). Dado el patrón sintáctico usado para términos en este estudio, tomamos en cuenta sólo nombres y adjetivos en ambos corpus porque son el tipo de palabras más usadas para la construcción de términos.

Gelbukh et al. (2010) y Kit & Liu (2008) sólo se enfocan en la extracción de candidatos unipa-

labra, por lo que únicamente ponderan las palabras que ocurren en ambos corpus. En nuestro experimento también consideramos las palabras que ocurren sólo en el corpus de dominio. En este sentido, las palabras con una frecuencia absoluta de al menos 1 se ponderan exclusivamente con la frecuencia de ocurrencia, como en 1, para el caso de las medidas razón de frecuencias relativas y diferencia de rangos. Para el caso de la medida razón *log-likelihood*, la ponderación únicamente se realiza con la frecuencia de la palabra para hacerla más compatible con la escala de ponderaciones de esta medida. En este trabajo, asumimos que el corpus de referencia es lo suficientemente grande para filtrar palabras no relevantes, por tanto las palabras que solo ocurren en el dominio tienen una probabilidad mayor de ser relevantes y su frecuencia refleja su importancia.

$$\text{Peso}\,(w_i) = 1 + \log_2\,(f_{w_i}) \qquad (1)$$

Consideramos que mientras más grande sea el corpus de referencia, mayor *exhaustividad*[2] tendrá de palabras de clase abierta de uso general, así como una probabilidad mayor de que ocurran términos de especialidad por lo menos una vez (el corpus de referencia fue recolectado de un periódico online donde se publican noticias respecto a ciencia y tecnología, así como otros rubros de información), por lo que, al igual que Drouin (2003), consideramos es lo suficientemente heterogéneo para contribuir a lograr una precisión más alta en la asignación de relevancia.

En resumen, como Ananiadou (1994) lo señala, la terminología se interesa en formas de palabra que ocurren con alta y baja frecuencia, o también aquellas en un rango medio, es decir, en todas las unidades que podrían ser términos en una colección de textos particular. Por su parte, Vivaldi et al. (2001) mencionan que incluso cuando se analizan corpus grandes, existe siempre la posibilidad de encontrar un término que sólo ocurra una vez. Por tanto, si esto representa o no un problema depende de la meta de la extracción. Por ejemplo, si solo deseamos caracterizar un documento, podríamos esperar que un término representativo ocurra muchas veces en el documento. Sin embargo, si deseamos hacer un mapa conceptual del texto, serían necesarios todos los términos.

[2] En el contexto de la recuperación de información, Baeza & Rivera (2011) describen la *exhaustividad* de un documento como la cobertura que proporciona para los temas principales del documento. Así, si agregamos nuevo vocabulario a un documento, la *exhaustividad* de la descripción del documento se incrementa.

4.5 Relevancia de términos candidatos

La relevancia de términos candidatos multipalabra se calcula como la suma de los pesos individuales (*termhood* unipalabra) de las palabras que forman el candidato.

Formalmente, si un sintagma nominal candidato s tiene una longitud de n palabras, $w_1 w_2 \ldots w_n$, donde $n > 1$, entonces la relevancia del candidato s es la suma de las ponderaciones individuales w_i de las palabras presentes en el sintagma:

$$\text{Termhood}\,(s) = \sum_{i=1}^{n} \text{termhood}\,(w_i) \qquad (2)$$

5 Resultados

Esta sección presenta los resultados de nuestro experimento con un conjunto de 200,000 tokens de un corpus extraído del sitio Web MedlinePlus en español.

5.1 Fuentes de información textual

Corpus de dominio

La fuente de información textual se constituye de un conjunto de documentos del dominio médico, básicamente enfermedades del cuerpo humano y temas relacionados (cirugías, tratamientos, etc.). Estos documentos se recolectaron del sitio Web MedlinePlus en español. MedlinePlus es un sitio que se enfoca en proporcionar información respecto a enfermedades, tratamientos y condiciones.

El tamaño del corpus es de 1.2 millones de tokens, pero realizamos nuestro experimento con un subconjunto de 200,000 tokens que refieren a enfermedades, tratamientos, etc., exclusivamente de los ojos. Decidimos restringir el corpus para ser capaces de determinar manualmente vía el uso de diccionarios y otros recursos confiables sobre el tema, el número de términos verdaderos presentes relacionados estrechamente con la temática del subcorpus y contar con una evaluación preliminar del desempeño de cada medida. Por último, seleccionamos un dominio médico debido a la disponibilidad de recursos textuales en formato digital.

Corpus de referencia

Con la meta de asignar relevancia a las palabras del dominio por medio un enfoque de contraste entre corpus, se recolectó automáticamente

un corpus de referencia de un periódico[3] online con artículos de noticias de todo el año 2014 (el tamaño del corpus es de aproximadamente de 5 millones de tokens).

Para construir el corpus se recolectaron los URLs del sitio con el módulo de Python BeautifulSoup[4]. Después, este conjunto de URLs se introdujo a la plataforma Sketch Engine[5] para recolectar automáticamente la información textual de cada página Web.

5.2 Otros recursos

Lenguajes de programación y otras herramientas

El lenguaje de programación usado para automatizar todas las tareas requeridas fue Python versión 3.4, así como el módulo NLTK version 3.0 Bird et al. (2009). Adicionalmente, el etiquetador morfosintáctico usado en este experimento fue FreeLing.

6 Análisis de resultados

En este experimento comparamos las medidas razón *log-likelihood* implementada por Gelbukh et al. (2010), la diferencia de rangos aplicada por Kit & Liu (2008), la razón de frecuencia relativa (Manning & Schütze, 1999) y la aproximación a la distribución binomial mediante el uso de la distribución normal estándar (Drouin, 2003) para medir *termhood* en palabras simples. Para las tres primeras medidas se aplica la reducción de ruido mencionada en la sección 4.3, así como la ponderación para las palabras que ocurren sólo en el dominio por medio de la frecuencia de ocurrencia. Para el caso de la medida usada en TermoStat[6] se utilizó dicha herramienta desde el sitio Web con el mismo subcorpus de dominio, sin embargo, dos factores que podrían sesgar los resultados para el caso de esta medida son el etiquetador morfosintáctico (el etiquetador usado por el sistema es TreeTagger) y el corpus de referencia, lo que afecta la comparación directa de los resultados con las tres medidas consideradas en este trabajo. Con todo, dada la cobertura altamente similar que se obtuvo del vocabulario que ocurre en ambos corpus (vocabulario común) se decidió considerarla en la comparación. Además, con el objetivo de hacer más comparables los resultados se lematizó con FreeLing el conjunto de palabras y los candidatos a término obtenidos con Termostat.

[3] http://www.lajornada.com.mx
[4] http://www.crummy.com/software/BeautifulSoup
[5] https://the.sketchengine.co.uk
[6] http://termostat.ling.umontreal.ca

Con la finalidad de comparar resultados para lograr un equilibrio entre precisión y cobertura, proponemos considerar los siguientes factores y algunas de sus combinaciones: candidatos sólo con vocabulario común (es decir, palabras que ocurren en ambos corpus), candidatos con vocabulario común y no común (incluyendo o no candidatos que ocurren una sola vez en corpus de dominio), filtro de adjetivos no relevantes de dominio y, finalmente, filtro de adjetivos no relevantes del corpus de referencia.

Vocabulario y términos del corpus

En el subcorpus analizado existen, en total, 1,842 palabras estrechamente relacionadas con el dominio y que participan en la construcción de la terminología implícita. De este subconjunto de palabras se derivan 2,253 términos unipalabra y multipalabra que cumplen con el patrón: `<NC><AQ>*`.

Ponderación del vocabulario común

El subcorpus de dominio tiene un total de 2,978 palabras que ocurren también en el corpus de referencia (vocabulario común). Estas palabras fueron ponderadas con cada una de las medidas comparadas en este trabajo: razón de frecuencia relativa (RFR), razón *log-likelihood* (RLL) y diferencia de rangos (DR). Como se mencionó en líneas anteriores, dado que el sistema Termostat considera otro corpus de referencia, el subconjunto de palabras ponderadas es diferente, en este caso específico corresponde a 1,696 palabras, por ello las celdas de los cuadros 3 y 4 en umbrales mayores que 2000 palabras no contienen datos. Estos cuadros muestran los niveles de precisión y cobertura del vocabulario común por umbrales de 500 palabras (ordenadas descendentemente por *termhood*). Así, esperaríamos que las palabras más relacionadas con el dominio se concentraran en las primeras posiciones. Por ejemplo, del primer subconjunto de 500 palabras, la cobertura obtenida por RFR, DR, TS y RLL es de 18.5 %, 18 %, 17.2 % y 14.4 %, respectivamente. Por otro lado, la precisión es de 68 %, 66.4 %, 63.2 % y 53.2 % para RFR, DR, TS y RLL, respectivamente. La cobertura total considerando sólo el vocabulario común para RFR, DR y RLL es del 43.3 % y 44.5 % para el caso de TS. Cabe señalar aquí que estos niveles de cobertura altamente similares no implican que los dos subconjuntos de palabras sean iguales, esto se debe a que consideran dos corpus de referencia diferentes.

Palabras	RLL	DF	RFR	TS
500	14.4	18	18.5	17.2
1000	24.4	29.2	30.7	30.7
1500	32.1	35.6	38.1	41.9
2000	39.7	39.2	41.9	44.5
2500	42.4	42	43.4	
3000	43.3	43.3	43.4	

Cuadro 3: Cobertura del vocabulario común.

Palabras	RLL	DF	RFR	TS
500	53.2	66.4	68	63.2
1000	45	53.8	56.6	56.5
1500	39.5	43.7	46.7	51.4
2000	36.6	36.1	38.6	41
2500	31.2	31	32	
3000	26.6	26.6	26.7	

Cuadro 4: Precisión del vocabulario común.

Ponderación del vocabulario común y no común

En este trabajo consideramos también la ponderación del vocabulario que ocurre sólo en el dominio mediante el uso de la frecuencia de ocurrencia. Por tanto asumimos que a mayor ocurrencia, mayor relevancia para el dominio. De los cuadros 5 y 6 se puede observar que se logra un aumento del 13.3 % en la cobertura del vocabulario relevante al dominio para el caso de la medida DF en los primeros 1000 candidatos, e incrementos menores para las medidas RFR y RLL. Por otro lado, la precisión más alta para las primeras 1000 palabras mejor ponderadas se obtiene con la medida DF, que alcanza un 78.3 %.

Palabras	RLL	DF	RFR	TS
500	14.7	22.1	18.7	17.2
1000	25.1	42.5	32.1	30.7
1500	35.2	62.4	42.3	41.9
2000	45.3	78.4	53.1	44.5
2500	54.6	86.4	64.1	
3000	67.5	91.6	77.9	
3500	82.4	95.4	97.2	
4000	97.3	97.1	97.9	
4500	97.8	97.8	97.9	

Cuadro 5: Cobertura del vocabulario común y no común.

Extracción y ponderación de términos sólo con vocabulario común

En los siguientes cuadros se muestran los datos de precisión y cobertura para los candidatos a término donde ocurren sólo palabras del vocabulario común, que constituyen el 43.3 % del vocabulario relevante al dominio para el caso de las medidas DR, RFR y RLL. La comparación de los datos en los cuadros 7 y 9 muestran el incremento en cobertura después de agregar los adjetivos del corpus de referencia a los extraídos del dominio mediante heurísticas lingüísticas con el objetivo de reducir ruido en los resultados. Por ejemplo, para el caso de los primeros 1000 candidatos, RLL, DF y RFR tienen un incremento por encima del 4 % en cobertura, esto debido a que se eliminan más adjetivos estrechamente relacionados con el contexto y que aparecen como modificadores de términos verdaderos (por ejemplo, *conjuntivitis rara*). Con respecto a la precisión, de los cuadros 8 y 10 se observa que hay un incremento en los primeros 1000 candidatos de 11.6 %, 11.5 %, 9.6 % para RLL, RFR y DR, respectivamente, donde las precisiones más altas y similares se encuentran en DR y RFR. Para el caso del sistema Termostat se obtuvo un conjunto de 2,695 candidatos unipalabra y multipalabra, por ello las celdas con umbrales mayores a 3000 candidatos no contienen información. La cobertura global con la reducción de ruido es de un 56.3 % sólo con el vocabulario común y después de la reducción de ruido en términos de los adjetivos tanto del corpus de dominio como los del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	54	81.6	68.8	63.2
1000	46.2	78.3	59.1	56.5
1500	43.2	76.6	51.9	51.4
2000	41.8	72.2	49.0	41.0
2500	40.2	63.7	47.2	
3000	41.4	56.2	47.8	
3500	43.4	50.2	51.2	
4000	44.8	44.7	45.1	
4500	40.0	40.0	40.1	

Cuadro 6: Precisión del vocabulario común y no común.

Palabras	RLL	DF	RFR	TS
500	13.9	14.3	14.6	7.4
1000	24.3	26.9	26.1	12.8
1500	33.4	38.3	37.9	16.4
2000	41.7	46.7	46.4	16.6
2500	49.7	51.4	52.3	16.6
3000	55.3	54.8	56.6	16.6
3500	58.0	57.2	58.1	
4000	58.5	58.5	58.5	

Cuadro 7: Cobertura en la extracción de términos sin filtrar adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	62.6	64.6	65.6	33.2
1000	54.7	60.7	58.7	28.9
1500	50.2	57.5	56.9	24.6
2000	47	52.7	52.3	18.7
2500	44.8	46.3	47.1	14.9
3000	41.5	41.1	42.5	12.4
3500	37.3	36.8	37.4	
4000	33.0	33.0	33.0	

Cuadro 8: Precisión en la extracción de términos sin filtrar adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	16.5	17.0	16.4	7.4
1000	29.4	31.2	31.2	12.8
1500	39.1	41.9	42.7	16.4
2000	47.6	48.1	49.9	16.6
2500	53.2	51.5	54.4	16.6
3000	55.7	54.8	55.8	16.6
3500	56.3	56.3	56.3	
4000	56.3	56.3	56.3	

Cuadro 9: Cobertura en la extracción de términos con filtro de adjetivos del corpus de referencia.

Extracción y ponderación de términos con vocabulario común y no común

Si consideramos el vocabulario común y no común, así como los candidatos con frecuencia mayor o igual que 1 y que solo ocurren en el corpus de dominio, se observa un incremento en cobertura de 5.2 %, 4.4 % y 4.2 % para RLL, DR y RFR en los primeros 1000 candidatos después de la reducción de ruido (véase cuadros 11 y 13). Los principales cambios en cobertura se presentan en los umbrales mayores que 2000. Con respecto a la precisión, después de la reducción de ruido se obtiene un incremento del 11.7 %, 9.9 % y 9.3 % para el caso de RLL, DR y RFR en los primeros 1000 candidatos, donde las precisiones más altas corresponden a RFR con un 72.7 % y DR con un 70.5 %.

Palabras	RLL	DF	RFR	TS
500	74.2	76.4	74	33.2
1000	66.3	70.3	70.2	28.9
1500	58.8	63.0	64.1	24.6
2000	53.6	54.2	56.2	18.7
2500	48.0	46.4	49.0	14.9
3000	41.8	41.1	41.9	12.4
3500	36.2	36.2	36.2	
4000	31.7	31.7	31.7	

Cuadro 10: Precisión en la extracción de términos con filtro de adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	13.9	14.3	15.9	7.4
1000	24.3	26.9	28.1	12.8
1500	33.4	38.4	39.2	16.4
2000	41.9	51.4	49.3	16.6
2500	49.9	65.3	57.3	16.6
3000	57.4	78.8	65.2	16.6
3500	66.6	83.1	74.1	
4000	76.9	86.5	85.4	
4500	89.9	88.8	90.1	
5000	90.1	90.1	90.1	

Cuadro 11: Cobertura en la extracción de términos sin filtro de adjetivos de corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	62.6	64.4	71.6	33.2
1000	54.7	60.6	63.4	28.9
1500	50.2	57.7	58.9	24.6
2000	47.2	57.9	55.6	18.7
2500	45.0	58.8	51.7	14.9
3000	43.1	59.2	48.9	12.4
3500	42.9	53.5	47.7	
4000	43.3	48.7	48.1	
4500	45.0	44.5	45.1	
5000	40.6	40.6	40.6	

Cuadro 12: Precisión en la extracción de términos sin filtro de adjetivos de corpus de referencia.

Por último, si eliminamos los candidatos que sólo ocurren en el corpus de dominio y con frecuencia igual a 1, la cobertura global es de un 73 % después de la reducción de ruido y la precisión para los primeros 1000 candidatos se mantiene prácticamente sin cambios respecto a los resultados incluyendo este subconjunto.

Palabras	RLL	DF	RFR	TS
500	16.5	17.0	17.5	7.4
1000	29.5	31.3	32.3	12.8
1500	39.2	43.1	44.8	16.4
2000	47.8	57.3	53.9	16.6
2500	55.6	70.8	62.8	16.6
3000	64.4	80.1	71.6	16.6
3500	75.5	83.3	82.9	
4000	87.6	86.3	87.8	
4500	87.8	87.8	87.8	
5000	87.8	87.8	87.8	

Cuadro 13: Cobertura en la extracción de términos con filtro de adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	74.2	76.4	79	33.2
1000	66.4	70.5	72.7	28.9
1500	58.9	64.7	67.3	24.6
2000	53.9	64.5	60.7	18.7
2500	50.1	63.8	56.6	14.9
3000	48.4	60.1	53.8	12.4
3500	48.6	53.6	53.3	
4000	49.4	48.6	49.5	
4500	44.0	44.0	44.0	
5000	39.6	39.6	39.6	

Cuadro 14: Precisión en la extracción de términos con filtro de adjetivos del corpus de referencia.

7 Conclusiones

En este trabajo hemos presentado una metodología para identificar y extraer términos unipalabra y multipalabra, reconocibles en un corpus de dominio especializado. Inicialmente, para asignar relevancia a palabras simples comparamos cuatro medidas diferentes e implementamos algunas heurísticas linguísticas, para reducir *ruido* en los resultados. De las cuatro medidas comparadas, la diferencia de rangos y la razón de frecuencia relativa fueron las que lograron los mejores resultados en términos de precisión y cobertura. Además, la estrategia para reducir ruido en los resultados, que consiste en considerar heurísticas de corte lingüístico para obtener adjetivos que con frecuencia no participan en la construcción de términos, derivó en buenos resultados al permitir aumentar precisión sin dañar significativamente la cobertura. Por otro lado, la propuesta de construir *termhood* multipalabra a partir del *termhood* de los elementos constituyentes proporcionó buenos resultados ya que un elemento con una ponderación individual alta contribuirá a incrementar el *termhood* de cualquier sintagma donde se encuentre presente.

Por tanto, el enfoque de comparación de corpus resultó útil para asignar relevancia a palabras de un dominio ya que asumimos que las palabras estrechamente relacionadas con el dominio analizado tendrán una mayor probabilidad de ocurrencia en el dominio que en un corpus de otro dominio diferente o de lengua general, lo que generará un *termhood* alto.

Actualmente existen muchas fuentes de información textual disponibles en la Web en varias lenguas, lo que facilita la obtención de documentos que sean útiles para implementar enfoques de comparación de corpus. En este sentido, el empleo de noticias de periódicos electrónicos es sumamente valioso, pues contienen información sobre muchos temas: cultura, política, ciencia, tecnología, etc., lo que favorece la heterogeneidad y exhaustividad del recurso textual, mejorando con ello la precisión al momento de implementar enfoques contrastivos.

Finalmente, resulta de gran interés probar la metodología propuesta en otros corpus de dominio para explorar la estabilidad de los resultados, lo que constituye parte de nuestro trabajo futuro.

Agradecimentos

Este trabajo ha sido patrocinado por la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), del Gobierno de Chile. Números de proyectos: 3140332 y 11130565.

Referencias

Acosta, Olga, Cesar Antonio Aguilar & Gerardo Sierra. 2013. Using relational adjectives for extracting hyponyms from medical texts. En Antonio Lieto & Marco Cruciani (eds.), *Proceedings of the First International Workshop on Artificial Intelligence and Cognition*, vol. 1100 CEUR Workshop Proceedings, 33–44.

Ahmad, Khurshid, Lee Gillam & Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). En *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland.

Ananiadou, Sophia. 1994. A methodology for automatic term recognition. En *Proceedings of the 15th Conference on Computational Linguistics - Volume 2* COLING '94, 1034–1038. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ananiadou, Sophia & John Mcnaught. 2005. *Text mining for biology and biomedicine*. Norwood, MA, USA: Artech House, Inc.

Aubin, Sophie & Thierry Hamon. 2006. Improving term extraction with terminological resources. En Tapio Salakoski, Filip Ginter, Sampo Pyysalo & Tapio Pahikkala (eds.), *Advances in Natural Language Processing*, vol. 4139 Lecture Notes in Computer Science, 380–387. Springer Berlin Heidelberg.

Baeza, Ricardo & Berthier Rivera. 2011. *Modern information retrieval*. Addison Wesley.

Barrón-Cedeño, Alberto, Gerardo Sierra, Patrick Drouin & Sophia Ananiadou. 2009. An improved automatic term recognition method for

spanish. En Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 5449 Lecture Notes in Computer Science, 125–136. Springer Berlin Heidelberg.

Basili, Roberto, Gianluca De Rossi & Maria Pazienza. 1997. Inducing terminology for lexical acquisition. En C. Cardie & R. Weischedel (eds.), *Proceeding of EMNLP 97 Conference*, 125–133.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with python*. O'Reilly.

Bourigault, Didier, Isabelle Gonzalez-Mullier & Cécile Gros. 1996. LEXTER, a natural language processing tool for terminology extraction. En Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström & Catalina Röjder Papmehl (eds.), *Proceedings of the 7th EURALEX International Congress*, 771–779. Göteborg, Sweden: Novum Grafiska AB.

Buitelaar, Paul, Philipp Cimiano & Bernardo Magnini. 2005. *Ontology learning from text: Methods, evaluation and applications*, vol. 123 Frontiers in Artificial Intelligence and Applications Series. Amsterdam: IOS Press.

Carreras, Xavier, Isaac Chao, Lluís Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. En *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, European Language Resources Association (ELRA).

Chung, Teresa. 2003. A corpus comparison approach for terminology extraction. *Terminology* 9(26). 221—-246.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16(1). 22–29.

Collier, Nigel, Chikashi Nobata & Jun-ichi Tsujii. 2000. Extracting the names of genes and gene products with a hidden markov model. En *Proceedings of the 18th Conference on Computational Linguistics - Volume 1* COLING '00, 201–207. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daille, Béatrice, Éric Gaussier & Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. En *Proceedings of the 15th Conference on Computational Linguistics - Volume 1* COLING '94, 515–521. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daille, Béatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology. En Judith L. Klavans & Philip Resnik (eds.), *The balancing act: Combining symbolic and statistical approaches to language*, 49–66. MIT Press.

Daille, Béatrice, Benoît Habert, Christian Jacquemin & Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology* 3(2). 197–257.

Demonte, Violeta. 1999. El adjetivo. clases y usos. la posición del adjetivo en el sintagma nominal. En Ignacio Bosque & Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, 129–215. Espasa.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1). 99–115.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19(1). 61–74.

Enguehard, Chantal & Laurent Pantera. 1995. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics* 2(1). 27–32.

Evans, David A. & Robert G. Lefferts. 1995. CLARIT-TREC experiments. En *Proceedings of the Second Conference on Text Retrieval Conference* TREC-2, 385–395. Elmsford, NY, USA: Pergamon Press, Inc.

Frantzi, Katerina, Sophia Ananiadou & Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries* 3(2). 115–130.

Fábregas, Antonio. 2007. The internal syntactic structure of relational adjectives. *Probus* 19(1). 1–36.

Gelbukh, Alexander, Grigori Sidorov, Eduardo Lavin-Villa & Liliana Chanona-Hernandez. 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. En ChristinaJ. Hopfe, Yacine Rezgui, Elisabeth Métais, Alun Preece & Haijiang Li (eds.), *Natural Language Processing and Information Systems*, vol. 6177 Lecture Notes in Computer Science, 248–255. Springer Berlin Heidelberg.

Heid, Ulrich. 1998. A linguistic bootstrapping approach to the extraction of term candidates from germ text. *Terminology* 5(2). 161–181.

Jacquemin, Christian. 1997. Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.

Jacquemin, Christian. 2001. *Spotting and discovering terms through natural language processing*. Cambridge: MIT Press.

Justeson, John S. & Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1. 9–27.

Kageura, Kyo & Bin Umino. 1996. Methods of automatic term recognition: a review. *Terminology* 3(2). 259–289.

Kazama, Jun'ichi, Takaki Makino, Yoshihiro Ohta & Jun'ichi Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. En *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3* BioMed '02, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kit, Chunyu & Xiaoyue Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology* 14(2). 204–229.

Kockaert, Hendrik & Frieda Steurs. 2015. *Handbook of terminology*, vol. 1. John Benjamins.

Lopez, Patrice & Laurent Romary. 2010. HUMB: automatic key term extraction from scientific articles in GROBID. En *Proceedings of the 5th International Workshop on Semantic Evaluation* SemEval '10, 248–251. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lossio-Ventura, JuanAntonio, Clement Jonquet, Mathieu Roche & Maguelonne Teisseire. 2014. Yet another ranking function for automatic multiword term extraction. En Adam Przepiórkowski & Maciej Ogrodniczuk (eds.), *Advances in Natural Language Processing*, vol. 8686 Lecture Notes in Computer Science, 52–64. Springer International Publishing.

L'Homme, Marie-Claude. 2004. *La terminologie: príncipes et techniques*. Les Presses de l'Université de Montréal.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Marchis, Mihaela. 2010. *Relational adjectives at the syntax/morphology interface in Romanian and Spanish*: Institut für Linguistik/Anglistik, Universität Stuttgart. Tesis Doctoral.

Matsuo, Yutaka & Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13(1). 157–169.

Maynard, Diana & Sophia Ananiadou. 1999. Identifying contextual information for multiword term extraction. En Peter Sandrini (ed.), *Proceedings of the TKE '99 International Congress on Terminology and Knowledge Engineering*, 212—-221. Vienna, Austria.

Medelyan, Olena & Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. En *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* JCDL '06, 296–297. New York, NY, USA: ACM.

Medina, Alfonso, Gerardo Sierra, Gabriel Garduño, Carlos Méndez & Roberto Saldaña. 2004. CLI. an open linguistic corpus for engineering. En Guillermo de Ita, Olac Fuentes & Mauricio Osorio (eds.), *Memorias del IX Congreso Iberoamericano de Inteligencia Artificial IBERAMIA 2004*, 203—-208. Puebla, México.

Pantel, Patrick & Dekang Lin. 2001. A statistical corpus-based term extractor. En Eleni Stroulia & Stan Matwin (eds.), *Advances in Artificial Intelligence*, 36–46. Springer.

Pazienza, Maria Teresa. 1998. A domain-specific terminology-extraction system. *Terminology* 5(2). 183–201.

Pazienza, MariaTeresa, Marco Pennacchiotti & FabioMassimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. En Spiros Sirmakessis (ed.), *Knowledge Mining*, vol. 185 Studies in Fuzziness and Soft Computing, 255–279. Springer Berlin Heidelberg.

Poesio, Massimo. 2005. Domain modelling and nlp: Formal ontologies? lexica? or a bit of both? *Applied Ontologies* 1(1). 27–33.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of the International Conference on New Methods in Language Processing*, 44–49.

Spasić, Irena, Mark Greenwood, Alun Preece, Nick Francis & Glyn Elwyn. 2013. FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics* 4(1).

Vivaldi, Jordi. 2004. *Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas*. Barcelona: IULA-UPF. Tesis Doctoral.

Vivaldi, Jordi, Lluís Màrquez & Horacio Rodríguez. 2001. Improving term extraction by system combination using boosting. En Luc De Raedt & Peter Flach (eds.), *Machine Learning: ECML 2001*, vol. 2167 Lecture Notes in Computer Science, 515–526. Springer Berlin Heidelberg.

Vivaldi, Jorge & Horacio Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13(2). 225–248.

Vivanco, Verónica. 2006. *El español de la ciencia y la tecnología*. Madrid: Arco Libros.

Wermter, Joachim & Udo Hahn. 2005. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 843–850. Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Wong, Wilson. 2008. Determination of unithood and termhood for term recognition. En Min Song & Yi-Fang Brook Wu (eds.), *Handbook of research on text and web mining technologies*, 500–529. Hershey, New York!: IGI Global.

Yamamoto, Kaoru, Taku Kudo, Akihiko Konagaya & Yuji Matsumoto. 2003. Protein name tagging for biomedical annotation in text. En *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 65–72. Sapporo, Japan: Association for Computational Linguistics.

A Apéndice

Un pequeño conjunto de términos candidatos antes y después de reducir ruido.

Previo a la reducción de ruido

enfermedad/NC ocular/AQ alérgico/AQ severo/AQ
vaso/NC sanguíneo/AQ anormal/AQ
catarata/NC congénito/AQ hereditario/AQ
degeneración/NC macular/AQ húmedo/AQ
ganglio/NC linfático/AQ sensible/AQ
vaso/NC sanguíneo/AQ permeable/AQ
vaso/NC sanguíneo/AQ retiniano/AQ
resequedad/NC ocular/AQ serio/AQ
vaso/NC sanguíneo/AQ defectuoso/AQ
gota/NC lubricante/AQ ocular/AQ
degeneración/NC macular/AQ seco/AQ
gota/NC oftálmico/AQ antibiótico/AQ
nervio/NC óptico/AQ saludable/AQ
anteojo/NC protector/AQ
cirujano/NC oftalmológico/AQ
vaso/NC sanguíneo/AQ cerebral/AQ
secreción/NC ocular/AQ seco/AQ
irritación/NC ocular/AQ leve/AQ
anteojo/NC común/AQ
degeneración/NC macular/AQ senil/AQ
examen/NC ocular/AQ estándar/AQ
trastorno/NC ocular/AQ común/AQ
ganglio/NC linfático/AQ circundante/AQ
músculo/NC ocular/AQ externo/AQ
movimiento/NC ocular/AQ anormal/AQ
enfermedad/NC ocular/AQ específico/AQ
órgano/NC digestivo/AQ
degeneración/NC macular/AQ temprano/AQ
vaso/NC sanguíneo/AQ frágil/AQ
oclusión/NC arterial/AQ retiniano/AQ
ganglio/NC linfático/AQ cercano/AQ
característica/NC facial/AQ anormal/AQ
anteojo/NC oscuro/AQ
afección/NC ocular/AQ grave/AQ
degeneración/NC macular/AQ intermedio/AQ
lente/NC intraocular/AQ artificial/AQ
tejido/NC corneal/AQ subyacente/AQ
coágulo/NC sanguíneo/AQ frecuente/AQ
cristalino/NC intraocular/AQ artificial/AQ
ceguera/NC nocturno/AQ congénito/AQ
enfermedad/NC intestinal/AQ inflamatorio/AQ
enfermedad/NC inflamatorio/AQ intestinal/AQ
dolor/NC ocular/AQ severo/AQ
inflamación/NC ocular/AQ
reflejo/NC nervioso/AQ anormal/AQ
ciclosporina/NC líquido/AQ
examen/NC ocular/AQ completo/AQ
dolor/NC ocular/AQ excesivo/AQ
nervio/NC craneal/AQ
examen/NC ocular/AQ minucioso/AQ
parálisis/NC facial/AQ periférico/AQ idiopático/AQ
ganglio/NC linfático/AQ
ganglio/NC linfático/AQ justo/AQ
inclinación/NC palpebral/AQ anormal/AQ
trastorno/NC neurológico/AQ agudo/AQ
examen/NC oftálmico/AQ estándar/AQ
conteo/NC sanguíneo/AQ completo/AQ
vaso/NC sanguíneo/AQ débil/AQ
músculo/NC ocular/AQ
órgano/NC hueco/AQ
presión/NC sanguíneo/AQ normal/AQ
esteroide/NC oftálmico/AQ suave/AQ
molestia/NC gastrointestinal/AQ leve/AQ
nervio/NC óptico/AQ

Después de la reducción de ruido

vaso/NC sanguíneo/AQ permeable/AQ
vaso/NC sanguíneo/AQ retiniano/AQ

cirujano/NC oftalmológico/AQ
vaso/NC sanguíneo/AQ cerebral/AQ
degeneración/NC macular/AQ senil/AQ
ganglio/NC linfático/AQ circundante/AQ
órgano/NC digestivo/AQ
oclusión/NC arterial/AQ retiniano/AQ
lente/NC intraocular/AQ artificial/AQ
cristalino/NC intraocular/AQ artificial/AQ
enfermedad/NC intestinal/AQ inflamatorio/AQ
enfermedad/NC inflamatorio/AQ intestinal/AQ
nervio/NC craneal/AQ
parálisis/NC facial/AQ periférico/AQ idiopático/AQ
ganglio/NC linfático/AQ
examen/NC oftálmico/AQ estándar/AQ
nervio/NC óptico/AQ
inflamación/NC articular/AQ
vaso/NC sanguíneo/AQ
órgano/NC abdominal/AQ
enfermedad/NC vascular/AQ cerebral/AQ
nervio/NC facial/AQ
cirujano/NC experto/AQ
examen/NC oftalmológico/AQ estándar/AQ
anteojo/NC
cirujano/NC especialista/AQ
músculo/NC facial/AQ
quiasma/NC óptico/AQ
degeneración/NC macular/AQ
gota/NC oftálmico/AQ antimicótico/AQ
neuritis/NC óptico/AQ autoinmunitario/AQ
enfermedad/NC gastrointestinal/AQ
cirujano/NC
presión/NC sanguíneo/AQ diastólico/AQ
retinopatía/NC diabético/AQ
alergia/NC nasal/AQ
cáncer/NC sanguíneo/AQ
traumatismo/NC craneal/AQ
torrente/NC sanguíneo/AQ
medicamento/NC tópico/AQ
cirugía/NC facial/AQ
resonancia/NC magnético/AQ cerebral/AQ
célula/NC sanguíneo/AQ
enfermedad/NC diabético/AQ
enfermedad/NC digestivo/AQ
gota/NC oftálmico/AQ profiláctico/AQ
gota/NC oftálmico/AQ antinflamatorias/AQ
gota/NC oftálmico/AQ homeopático/AQ
órgano/NC corporal/AQ
trastorno/NC articular/AQ
imagen/NC visual/AQ
lente/NC oftálmico/AQ
conteo/NC sanguíneo/AQ
presión/NC sanguíneo/AQ
hipertensión/NC arterial/AQ
trastorno/NC digestivo/AQ
tic/NC facial/AQ
trastorno/NC visual/AQ
hinchazón/NC facial/AQ
enfermedad/NC intestinal/AQ
inflamación/NC viral/AQ
infección/NC gastrointestinal/AQ
neuropatía/NC óptico/AQ isquémico/AQ
coágulo/NC sanguíneo/AQ

Adjetivos obtenidos del corpus de dominio.

severo, común, cierto, probable, mejor, distinto, contagioso, posible, propio, bueno, conveniente, viejo, tratable, temprano, siguiente, susceptible, alto, cuidadoso, diverso, agresivo, mayor, pequeño, efectivo, bajo, importante, intenso, visible, presunto, reseco, diferente, útil, complejo, único, largo, gran, mismo, miope, borroso, corto, redondo, saludable, rojo, abundante, inconsciente, simple, atento, peligroso, principal, nuevo, suficiente, grande, excesivo, húmedo, obvio, necesario, suceptivas, disponible, caliente, molesto, menor, usual, grave, evitable, calmo, serio, fino, eficaz, solo, cercano, excelente, pendiente, activo, frecuente, rápido, transparente, profesional, national, opaco, difícil, brillante, seguro, leve, raro, frágil, mediano, lento, potente, fácil, resistente, amplio, delgado, fuerte, american, ocular, específico, numeroso, precoz, completo, doloroso, sensible, infecto, sano, profundo, particular, especial, lleno, genial, increíble, grueso, pegajoso, oscuro, próximo, ciego, habitual, diminuto, joven, hereditario, presente, terapéutico, sólido, breve, incierto, dudoso, aceptable, claro, professional, antecedente, normal, estricto, mortal, duro, valioso, esencial, múltiple, regular, muscular, eficiente, desastroso, dañino, sensitivo, prominente, frustrante, típico, costoso, significativo, reciente, accesible, notorio, espeso, quieto, diario, delicado, vulnerable, constante, medio, proclive, estrecho, verdadero, evidente, clásico, enfermo, seco, denso, fijo, igual, menudo, distante, despierto, derecho, gafo, total, extenso, vertical, incómodo, agudo, tremendo, impresionante, inevitable, flexible, capaz, contento, general, soñoliento, característico, peor, preciso, international, riesgoso, rosado, invisible, ligero, suave, frío, tóxico, invasivo, variable, existente, pálido, futuro, débil, propenso, responsable, alérgico, nítido, graso, antiguo, prematuro, anormal, ácido, áspero, anciano, blanco, doble, letárgico, amarillo, flojo, rígido, tardío, extraño, cómodo, junto, poderoso, asimétrico, viscoso, orgulloso, moderno, lens, triste, famoso, posterior, difuso, sencillo, esférico, correcto, interno, externo, congénito, máximo, corriente, degenerativo, listo, problemático, enorme, explorador, malo, extremo, afecto