Extracción de términos a través de la comparación entre corpus

Term extraction through the comparison between corpora

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

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

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

Edith Ramírez
Magíster en Procesamiento y Gestión de la Información
Pontificia Universidad Católica de Chile
Santiago de Chile
edithramirezll@gmail.com

Resumen: En este trabajo se desarrolla un método híbrido para la extracción de términos de un corpus especializado en español. En una primera fase se consideran los patrones lingüísticos de los términos más recurrentes en dominios especializados. En paralelo, se toma en cuenta un enfoque de comparación de corpus para calcular la relevancia de cada palabra (ing.: *termhood*) en el dominio tratado. Posteriormente, un conjunto de adjetivos no relevantes del corpus de referencia y del dominio se filtran de los candidatos para eliminar aquellos modificadores que están estrechamente ligados al contexto y que no forman parte de un término (*relevante*, *raro*, *fuerte*, etc.). En una segunda fase se configura la relevancia de los candidatos a término a partir del valor de cada uno de sus componentes. Nuestro método obtiene mejores resultados que el método *C-Value/NC-Value* aplicado al español (Barrón *et al.*, 2009), el cual hemos aplicado al mismo corpus de medicina.
Palabras clave: Término, extracción terminológica, termicidad, unicidad, ontología.

Abstract: This paper presents a hybrid method for the extraction of terms from a specialized Spanish corpus. In a first step we consider a set of linguistic patterns of more recurrent terms in specialized domains. In parallel, we take into account a comparison approach between corpora to estimate the relevance of each word in its knowledge domain (that is, its degree of *termhood*). Subsequently, a set of non-relevant adjectives taken from reference and domain corpora is filtered from candidates in order to eliminate those modifiers that are closely linked to the context and not part of a term (*relevant*, *rare*, *strong*, etc.). In a second phase, we configure the relevance of term candidates according to the value of each of its components. Our method obtains better results than the *C-value/NC-Value* method applied to Spanish (Barrón *et al.*, 2009).
Keywords: Term, term extraction, termhood, unithood, ontology.

1 Introducción

Una de las fases más importantes para el desarrollo de ontologías —en concreto aquellas que obtienen su información a partir de fuentes textuales especializadas— es la extracción de términos (Buitelaar, Cimiano y Magnini, 2005). Autores como Kageura y Umino (1996), Daille *et al.* (1996), Pazienza (1998), Jacquemin y Bourigault (2003), así como Vivaldi y Rodríguez (2007) definen al término como una etiqueta lingüística asociada a un concepto específico de dominio. Tomando en cuenta esta definición, Pazienza, Pennacchiotti y Zanzotto (2005) señalan que la implementación de métodos para la extracción automática de términos es esencial para crear cualquier ontología.

Existen al menos tres enfoques para desarrollar un método de extracción

terminológica. Por un lado, se han utilizado medidas estadísticas para la identificación y clasificación de candidatos a términos, ponderando sobre todo la asignación de una medida que indique su grado de relevancia dentro de un dominio de conocimiento, lo que se ha denominado en inglés *termhood*, así como la estabilidad de unidades sintagmáticas candidatas a término, en inglés, *unithood* (Kageura y Umino, 1996).

Por otro lado, existen también técnicas lingüísticas que permiten filtrar candidatos relevantes mediante la identificación de patrones sintácticos específicos (Heid, 1999). Finalmente, varios autores han planteado métodos híbridos que incorporan fases tanto lingüísticas como estadísticas, en aras de lograr una mayor exactitud respecto a la identificación de términos (Frantzi, Ananiadou y Tsujii, 1998; Kageura y Umino, 1996; Vivaldi y Rodríguez, 2007).

Situándonos en este último enfoque, proponemos aquí un método híbrido, el cual toma en cuenta el uso de un corpus de referencia grande, con el fin de establecer la relevancia de las palabras que ocurren en ambos corpus. Tales valores son generados a partir de razones de frecuencias relativas (Manning y Schütze, 1999). Vale decir aquí que el uso de un corpus de referencia nos parece un método eficaz para darle un valor relevante al vocabulario especializado reconocible en ambos corpus, considerando que la distribución de términos verdaderos y no verdaderos varía significativamente en diferentes tipos de texto (Kit y Liu, 2008). Así, con esta medida, una parte significativa del vocabulario que define un dominio de conocimiento tendrá una ponderación alta, y, en consecuencia, aquellos sintagmas candidatos donde se encuentren estos elementos tendrán una ponderación alta.

La distribución de nuestro trabajo es la siguiente: en la sección 2 planteamos una breve revisión del estado del arte sobre extracción terminológica. En la sección 3 analizamos el valor que tienen los adjetivos como unidades terminológicas, de tal suerte que este valor nos permite plantear una serie de heurísticas lingüísticas para la eliminación de adjetivos no relevantes asociados a nombres, en aras de reconocer y extraer buenos candidatos. En la sección 4 describimos nuestra metodología de extracción. En la sección 5 exponemos nuestros resultados y, finalmente, en la sección 6 delineamos nuestras conclusiones.

2 Estado del arte sobre extracción terminológica

Desde sus comienzos, la extracción de información ha estado estrechamente involucrada con la identificación de términos en corpus. De acuerdo con Wüster (1979), los nombres y los sintagmas nominales son las estructuras sintácticas canónicas para codificar términos. Tomando en cuenta este planteamiento, una buena parte de los experimentos para extraer términos ponen énfasis en el uso de patrones léxico-sintácticos. Un buen ejemplo de esto son los experimentos realizados por Heid para el alemán (1998, 1999), en donde detalla una metodología que permite identificar secuencias de términos a través del uso de expresiones y gramáticas regulares.

Otra vía que ha sido considerada es el uso de métodos estadísticos para asignar el grado de relevancia de candidatos a términos (*termhood*), así como el grado de estabilidad sintagmática de candidatos multi-palabra (*unithood*). Dentro de este enfoque puede considerarse el trabajo pionero de Salton y Yang (1973), el extractor desarrollado por Dagan y Church (1994) llamado *Termight*, o el algoritmo *C-Value*/*NC-Value* desarrollado por Frantzi, Ananiadou y Tsujii (1998).

Finalmente, la mayor parte de los experimentos de extracción terminológica han optado por emplear métodos híbridos que permiten combinar criterios lingüísticos y estadísticos, obteniendo así mejores resultados. Una buena perspectiva sobre cómo operan estos métodos la dan Kageura y Umino (1996), y más recientemente, Pazienza, Pennacchiotti, y Zanzotto (2005). Por su parte, Cabré, Estopà y Vivaldi (2001), evalúan el desempeño de 12 extractores diseñados para diferentes lenguas. Finalmente, Jacquemin y Bourigault (2003) ofrecen una panorámica resumida y precisa sobre esta clase de métodos.

3 El método híbrido C-Value/NC-Value aplicado al español

Un método híbrido que ha logrado obtener buenos resultados al extraer términos en inglés es el que desarrollan Frantzi, Ananiadou y Tsujii (1998), el cual consiste en la implementación de un algoritmo llamado *C-Value*/*NC-Value*.

Una aplicación de este algoritmo a textos en español la han llevado a cabo Barrón *et al.* (2009), quienes proponen una serie de modificaciones para mejorar el desempeño del algoritmo en general, así como adaptaciones para su aplicación a corpus en español. Básicamente, tales modificaciones consisten en:

- La eliminación de palabras no relevantes de los candidatos, en lugar de borrar el candidato completo de la lista (proceso de eliminación selectiva).
- La consideración de candidatos de longitud 1.
- La lematización del corpus de entrada para evitar la dispersión de frecuencias de ocurrencia por variaciones de un mismo candidato.
- La consideración de límites flexibles para las ventanas de contexto en la etapa NC-Value.

4 Adjetivos como unidades con valor terminológico

Si bien el experimento planteado por Barrón y otros representa un avance en la implementación del algoritmo *C-Value/NC-Value* como un método híbrido idóneo para la extracción de términos, los resultados que presentan en precisión y cobertura no logran un balance relevante. La evaluación del extractor, con los patrones lingüísticos más comunes (filtro lingüístico cerrado) y el filtrado de palabras no relevantes en candidatos, logran una precisión del 31%, en tanto que su cobertura alcanza un 50% (Barrón et al., 2009).

Desde nuestra perspectiva, uno de los factores que afecta este bajo rendimiento en los valores de precisión y cobertura es justo el uso de una *stop-list* elaborada subjetivamente. Aunado a la consideración de palabras funcionales como no relevantes, también categorías gramaticales como nombres y adjetivos contienen una gran proporción de elementos que no son relevantes en dominios especializados. Consideramos que la construcción automática de una lista de palabras no relevantes extraída del dominio es una tarea importante que se debe enfocar en las metodologías para la extracción de términos. Los elementos de esta lista podrían lograr una buena depuración de candidatos.

Atendiendo en particular el rol de los adjetivos como unidades terminológicas, Acosta, Aguilar y Sierra (2013) han analizado su función para describir los atributos asociados a un nombre, dando lugar a una categorización del referente expresado por dicho nombre dentro de un dominio conceptual específico.

Este proceso de categorización puede ser observado a detalle cuando se analiza el comportamiento semántico que siguen los adjetivos calificativos y relacionales al ligarse a nombres. El distinguir tales comportamientos puede ser sumamente útil para localizar buenos candidatos a términos dentro de un corpus.

4.1 Obtención de adjetivos no relevantes

Basándonos en la descripción que da Demonte (1999) sobre los adjetivos en español, estos son unidades sintácticas cuya función es la de modificar el significado de un nombre, asociándolo con uno o varios atributos. Entre los atributos que pueden ser introducidos por un adjetivo, están aquellos que caracterizan una propiedad física: color, tamaño, peso, forma, etc., identificados como *calificativos*: *grave, amarillo*; en contraste con aquellos que señalan rasgos que adscriben al nombre dentro de una clase específica, los cuales son denominados como *relacionales*. Ejemplos de estos adjetivos son: *muscular, venérea* y otros similares. Tomando en cuenta este comportamiento semántico de los adjetivos relacionales, resulta pertinente verlos como unidades constitutivas de términos. Un área que hace un enorme uso de estos adjetivos para construir términos con un alto grado de especialización es precisamente la medicina.

4.2 Particularidades lingüísticas de los adjetivos no relevantes

Al profundizar en el comportamiento semántico de los adjetivos, se observan otros rasgos distintivos importantes. De acuerdo con Demonte (1999), si se atiende su estructura interna, se puede hacer una distinción entre adjetivos permanentes y episódicos. Los primeros describen atributos estables y permanentes de una entidad (*psicópata*, *egocéntrico*, *apto*, etc.). Por su parte, los segundos dan cuenta de propiedades que están limitadas a restricciones temporales o espaciales (*harto*, *limpio*, *seco*, etc.).

Ligado a estas diferencias, en el español existen algunas particularidades sintácticas que influyen directamente en el significado de los

adjetivos. Así, los adjetivos permanentes tienden a relacionarse con el verbo *ser*, mientras que los episódicos se asocian al verbo *estar*. Esta distinción es clave para implementar una heurística que permita realizar una eliminación selectiva de estos últimos en sintagmas nominales candidatos.

Por otro lado, Demonte (1999) señala un conjunto de heurísticas sintácticas que permiten discernir entre adjetivos relacionales y calificativos. Por ejemplo, si un adjetivo va precedido de un adverbio, entonces, la probabilidad de que se trate de un adjetivo calificativo es alta: *muy alto*, *bastante raro*, etc.).

4.3 Interpretación composicional de nombres y adjetivos relacionales

De acuerdo con Demonte (1999), los adjetivos relacionales y los nombres que modifican tienen una interpretación composicional, es decir, en el caso de la construcción *enfermedad cardiovascular*, el adjetivo relacional *cardiovascular* vincula los rasgos o propiedades del *corazón* y *vasos sanguíneos* al nombre *enfermedad*. Esta situación complica el cálculo de medidas de estabilidad sintagmática (*unithood*), ya que tanto los nombres como los adjetivos relacionales pueden relacionarse con múltiples núcleos nominales o adjetivos.

5 Metodología

En este trabajo proponemos una metodología para extraer términos de un corpus de dominio especializado con etiquetado POS. La entrada al algoritmo debe ser el lema y la etiqueta POS correspondiente a la palabra.

5.1 Etiquetado de partes de la oración

El etiquetado de partes de la oración es el proceso de asignar una categoría gramatical a cada palabra en un corpus. Estas etiquetas permiten la configuración de un conjunto de patrones que filtren aquellos candidatos a término que tengan una mayor probabilidad de ser términos verdaderos. El etiquetador usado en nuestro trabajo es *FreeLing* (Carreras *et al.*, 2004) que está basado en las etiquetas del grupo EAGLES.

5.2 Chunking

Chunking (o análisis sintáctico superficial) es un proceso que consiste en dividir una oración en segmentos correlacionados sintácticamente llamados *chunks*, o sintagmas base. Una gramática para el análisis sintáctico superficial es un conjunto de reglas que indica cómo se deben agrupar las oraciones.

En este trabajo nos enfocamos en la extracción de términos con una estructura sintáctica específica. En este sentido, Vivaldi y Rodríguez (2007) presentan datos respecto a la estructura sintáctica de 2,145 términos reales en español derivados del corpus Técnico del IULA. Los datos muestran que un 48% de los términos son nombres únicos (por ejemplo, *síndrome*) y un 45% un nombre más adjetivos (por ejemplo, *vaso sanguíneo retiniano*). El 7% restante son términos que contienen una preposición, por ejemplo, *síndrome de Down*, donde la preposición *de* es la más recurrente. Respecto a esta última estructura sintáctica, en términos semánticos, la preposición *de* es la menos informativa de todas porque puede representar una gran cantidad de relaciones (Daille *et al*., 1996). Finalmente, los compuestos nominales en lenguas romance (por lo menos en español y rumano) son de difícil construcción (Marchis, 2010), lo que en inglés es muy recurrente y genera una gran cantidad de términos, por ejemplo, *kidney disease*. Por tanto, dada la escasa representación de frases con preposición *de* en la estructura sintáctica de términos y sus múltiples usos en el discurso especializado, así como la poca productividad de compuestos nominales en español, decidimos enfocar nuestro experimento solo en el patrón sintáctico <NC><AQ>*.

Por otro lado, las expresiones regulares para extraer adjetivos no relevantes toman en cuenta lo mencionado en la sección 4.2: adjetivos precedidos por el verbo *estar* (FreeLing hace esta distinción en su etiquetado: VS para *ser* y VA para verbos auxiliares, entre ellos *estar*), y adverbios antes de un adjetivo:

<RG><AQ>
<VAE><AQ>

Donde RG, AQ y VAE (para distinguir el verbo *estar*, se amplió la etiqueta a tres caracteres), de acuerdo con el etiquetado de FreeLing, corresponden a adverbios, adjetivos y verbo *estar* respectivamente.

5.3 Eliminación de ruido en candidatos

Proponemos aquí la eliminación selectiva de adjetivos y nombres no relevantes de los sintagmas candidatos a término.

En el caso de adjetivos, consideramos un par de heurísticas planteadas por Demonte (1999) que dan en conjunto una precisión por encima del 90% de adjetivos no relevantes (*raro, importante, distante,* etc.). Aunado a los adjetivos extraídos del dominio tratado con estas heurísticas, se considera un conjunto de 1,068 adjetivos del corpus de referencia que han sido revisados manualmente y que puede ser considerada como una lista base. Así, la lista de adjetivos no relevantes se conforma de los adjetivos del corpus de referencia y los extraídos del dominio.

Para el caso de nombres, se considera en este proceso de eliminación selectiva un conjunto pequeño de nombres no relevantes: *caso, conjunto, subconjunto, tipo, subtipo, forma, parte, clase, subclase*. Si el sintagma candidato contiene alguno de estos elementos como núcleo nominal, se elimina totalmente de la lista de candidatos.

5.4 Cálculo de relevancia de palabras

Consideramos el enfoque de razón de frecuencia relativa (Manning y Schütze, 1999) como en (1) entre dos corpus para asignar relevancia a las palabras comunes en ambos: dominio y referencia. Para este fin, tomamos en cuenta solo nombres y adjetivos porque son las categorías que más se usan en la construcción de términos y que además contemplamos en nuestro filtro sintáctico:

$$Relevancia(w_i) = log_2\left(1 + \frac{\frac{^{x}dom}{^{N}dom}}{\frac{^{x}ref}{^{N}ref}}\right) \quad (1)$$

Donde ^{x}dom, ^{N}dom corresponden a frecuencia de ocurrencia de w_i en dominio y tamaño del vocabulario de nombres y adjetivos del dominio, respectivamente. De forma semejante, ^{x}ref, ^{N}ref corresponden a la frecuencia de ocurrencia de la palabra w_i y tamaño del vocabulario del corpus de referencia. La unidad sumada al cociente en la fórmula (1) tiene como finalidad evitar valores negativos. Sin embargo, si se restringe el cálculo de relevancia de w_i solo a aquellas palabras donde la frecuencia relativa en dominio sea mayor que la del corpus de referencia, este valor puede eliminarse. Por otro lado, el cálculo de relevancia de las palabras que solo ocurren en el corpus de dominio se realiza considerando la formula (2):

$$Relevancia(w_i) = 1 + log_2(f_{w_i}) \quad (2)$$

Finalmente, asumimos que las palabras que ocurren solo en el corpus de dominio tienen una probabilidad alta de ser parte del vocabulario especializado debido a que el corpus de referencia con el que se contrasta es grande (4.5 millones de palabras). Así, suponemos que a mayor tamaño de corpus de referencia, mayor precisión de este conjunto de palabras.

5.5 Cálculo de relevancia de candidatos a término

El cálculo de relevancia al dominio de cada sintagma candidato a término se realiza mediante la suma de las relevancias de cada palabra presente en el candidato. Por ejemplo, si un sintagma nominal candidato *sn* tiene una longitud de n palabras, $w_1w_2…w_n$, entonces la relevancia del candidato *sn* es la suma de los pesos de todas las palabras w_i:

$$Relevancia(sn) = \sum_{i=1}^{n} weigth(w_i) \quad (3)$$

6 Resultados

A continuación, exponemos los resultados que obtuvimos aplicando ambos métodos híbridos: la adaptación C-Value/NC-Value propuesta por Barrón *et al*. (2009) y nuestro método.

6.1 Fuentes de información textual

6.1.1 Corpus de dominio

Nuestro corpus de dominio 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 de *MedlinePlus* en español[1]. El corpus es un subconjunto de 200,000 palabras para el que se determinó el número de términos presentes manualmente. Este subcorpus es parte de un conjunto mayor de 1.2 millones de palabras. Seleccionamos un dominio médico por razones de disponibilidad de fuentes textuales en formato digital. Además, asumimos que esta selección no supone restricciones muy fuertes

[1] Véase: www.nlm.nih.gov/medlineplus/spanish.

para la generalización de resultados a otros dominios.

6.1.2 Corpus de referencia

Con el fin de asignar relevancia a las palabras por medio de la razón de frecuencia relativa, se recolectó un corpus de referencia de 4.5 millones de palabras de un periódico mexicano que mantiene una versión en línea[2]. Se recolectaron las noticias de todo el año 2014. En una primera fase, se obtuvieron los URLs con la librería de Python *BeautifulSoup*[3] estableciendo un nivel de navegación específico. Posteriormente, se utilizó este conjunto de URLs en la herramienta *WebBootCat* de *Sketch Engine*[4] (Kilgarriff *et al.*, 2004) para extraer la información textual de cada página.

La representación por categoría de información en el corpus de referencia se presenta en la tabla 1.

Categoría	No. de doctos.	%
Ciencia	24	0.4
Política	1865	29.3
Espectáculos	98	1.5
Deportes	515	8.1
Sociedad	416	6.5
Capital	424	6.7
Estados	449	7.1
Economía	658	10.4
Mundo	662	10.4
Cultura	137	2.2
Edito	316	5.0
Correo	318	5.0
Opinión	319	5.0
Pág. principal	155	2.4

Tabla 1: Estructura del corpus de referencia.

6.2 Otros recursos

El lenguaje de programación usado para automatizar todas las tareas requeridas fue Python, versión 3.4, así como el módulo *NLTK*[5] (Bird, Klein y Loper, 2009), en su versión 3.0. Por otro lado, el etiquetador de partes de la oración usado fue *FreeLing*, la versión incorporada en la herramienta *Sketch Engine*.

6.3 Análisis de resultados

Con nuestra metodología obtuvimos, sin especificar un umbral y sin eliminar ruido en los candidatos, una precisión del 37.9% y una cobertura de 86.6%. El algoritmo C-Value/NC-Value se aplicó utilizando un filtro lingüístico cerrado con el patrón lingüístico <NC><AQ>* y como lista de palabras no relevantes aquellas con frecuencia mayor o igual que 100 en el corpus de referencia, verificando manualmente que no formaran parte de términos del dominio tratado, de acuerdo con los criterios mencionados en Barrón *et al*., (2009). Además, consideramos como mínima frecuencia de ocurrencia el valor 1. Finalmente, para la fase NC-Value consideramos el 25% de los mejores candidatos de la etapa C-Value. Con la especificación de parámetros anterior se obtuvo una precisión del 42% y una cobertura del 82%.

Para el caso del método aquí propuesto, la fase de eliminación de ruido, solo considerando los adjetivos extraídos automáticamente del dominio y la lista de nombres mencionada en la sección 5.4, mejoró los resultados hasta alcanzar una precisión de 46.9% y una cobertura de 85.1%. La ampliación de la lista de adjetivos para considerar también los adjetivos del corpus de referencia resulta en una precisión del 53.6% con una cobertura del 84.6%. La tabla 2 resume los resultados obtenidos para varios umbrales de relevancia en términos de precisión para ambos métodos.

Cand	C-Value/ NC-Value	Compara-ción entre corpus
1000	51%	84%
2000	47%	61%
3000	42%	54%
4000	42%	51%
5000	40%	49%
6000	42%	48%

Tabla 2. Comparación de resultados.

Los datos de la tabla 2 muestran un mejor desempeño de nuestro método comparado con las modificaciones propuestas por Barrón *et al*., (2009) al algoritmo C-Value/NC-Value. Sin embargo, consideramos indispensable la

[2] *La Jornada*: www.jornada.unam.mx.
[3] Véase: www.crummy.com/software/BeautifulSoup.
[4] Véase: www.sketchengine.co.uk.
[5] Véase: www.nltk.org.

aplicación del mismo método a otros dominios para constatar su estabilidad.

7 Conclusiones

En este artículo hemos presentado una metodología híbrida para extraer términos de un corpus de dominio especializado. Creemos que una primera fase de cálculo de relevancia de palabras mediante la consideración de un corpus de referencia grande, sea de lengua general o de otro dominio diferente al tratado, es un método efectivo porque la distribución de términos verdaderos y aquellos que no lo son varía de forma significativa dependiendo del tipo de texto que se trate.

Aunado a lo anterior, en este trabajo enfatizamos la necesidad de construir automáticamente una lista de instancias de palabras que no participan en la construcción de términos (por ejemplo, adjetivos como *raro*, *relevante*, y nombres como *tipo*, *forma*, etc.), pero que se encuentran dentro de las categorías más comunes para construirlos (nombres y adjetivos). En este sentido, la investigación lingüística relacionada con adjetivos, así como su constatación empírica en grandes corpus en español, ha sido de gran ayuda para implementar estos hallazgos lingüísticos y con ellos mejorar el proceso de eliminación selectiva en aras de obtener resultados más precisos sin pérdida significativa de cobertura. La interpretación composicional de candidatos dificulta el trabajo de medidas enfocadas a calcular la estabilidad sintagmática de candidatos, es por ello que consideramos que esta propiedad sigue siendo un problema abierto y es necesario buscar nuevas formas de enfocarla.

Concretamente sobre el experimento realizado y la propiedad de *unithood*, de acuerdo con Vivaldi y Rodríguez (2007), la estabilidad sintagmática se puede derivar del filtro sintáctico, lo que entendemos como: si dos o más palabras ocurren juntas y cumplen con la restricción sintáctica, entonces están asociadas. Así, proponemos en un futuro experimento considerar la suma de la frecuencia de ocurrencia del sintagma nominal como un todo en la fórmula (3) de la sección 5.5.

Los enfoques de comparación de corpus como el propuesto son cada vez más factibles debido a que actualmente existen muchas fuentes de información textual disponibles en línea, por ejemplo, la información de periódicos que explotamos en este trabajo y que se pueden extraer de una forma simple y rápida explotando la tecnología actual. Además, estas fuentes regularmente ofrecen información variada en términos de política, cultura, ciencia, etc., lo cual aumenta la probabilidad de ocurrencia de términos especializados.

Actualmente estamos en proceso de recolección de una mayor cantidad de información sobre la categoría *ciencia* vía la aplicación de la misma metodología para noticias de otros años, esto con la finalidad de incrementar su representación en el corpus, lo que suponemos se traducirá en una mejora significativa de los resultados.

Reconocimiento

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

Bibliografía

Acosta, O., C. Aguilar, y G. Sierra. 2013. Using Relational Adjectives for Extracting Hyponyms from Medical Texts. En A. Lieto y M. Cruciani (eds.) *Proceedings of the First International Workshop on Artificial Intelligence and Cognition*, páginas 33-44. CEUR Workshop Proceedings, Turín.

Barrón, A., G. Sierra, P. Drouin y S. Ananiadou. 2009. An Improved Automatic Term Recognition Method for Spanish. En A. Gelbukh (ed.) *Proceedings of CICLing 2009*, *LNCS*, páginas 125-136. Berlín, Springer.

Bird, S., E. Klein y E. Loper. 2009. *Natural Language Processing with Python*. Sebastropol, Cal., USA, O'Reilly.

Buitelaar, P., P. Cimiano y B. Magnini. 2005. *Ontology learning from text*. Amsterdam, IOS Press.

Cabré, M. T., R. Estopà y J. Vivaldi. 2001. Automatic term detection: A review of current systems. En D. Bourigault, C. Jacquemin y M.-C. L'Homme (eds.) *Recent advances in computational terminology*, páginas 53-88. Amsterdam/Philadelphia, John Benjamins Publish.

Carreras, X., I. Chao, L. Padró y M. Padró. 2004. *FreeLing: An Open-Source Suite of*

Language Analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004*, páginas 239-242. ELRA Publications, Lisboa.

Dagan, I., y K. Church. 1994. Termight: Identifying and translating technical terminology. En *Proceedings of the Fourth Conference on Applied Natural Language Processing*, páginas 34-40. Stuttgart.

Daille, B., B. Habert, C. Jacquemin, y J. Royauté. 1996. Empirical Observation of Term Variations and Principles for their Description. *Terminology* 3(2): 197-257.

Demonte, V. 1999. El adjetivo. Clases y usos. La posición del adjetivo en el sintag-nominal. En I. Bosque y V. Demonte (eds.) *Gramática descriptiva de la lengua española*, Vol. 1, Cap. 3, páginas 129-215. Madrid, Espasa.

Frantzi, K., S. Ananiadou y J. Tsujii. 1998. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. En C. Nikolaou y C. Stephanidis (eds.) *Research and Advanced Technology for Digital Libraries*, *LNCS*, páginas 585-604. Berlín, Springer.

Heid, U. 1999. Extracting terminologically relevant collocations from German technical texts. En P. Sandrini (ed.) *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering*, páginas 241-255. Universität Innsbruck, Innsbruck, Austria.

Jacquemin, C. y D. Bourigault. 2003. Term Extraction and Automatic Indexing. En R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, páginas 599-615. Oxford, UK. Oxford University Press.

Kageura, K., y B. Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(2), 259-289.

Kilgarriff, A., P. Rychlý, P. Smrž y D. Tugwell. 2004. The Sketch Engine. En *Proceedings of the 11th EURALEX International Congress*, páginas 105-116. Lorient, Francia.

Kit, C., y X. Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison, *Terminology*, 14(2): 204-229.

Manning, C., y H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass. MIT Press.

Marchis, M. 2010. *Relational Adjectives at the Syntax/Morphology Interface in Romanian and Spanish*. Ph. D. Dissertation. Stuttgart, Institut für Linguistik, Universität Stuttgart.

Pazienza, M., M. Pennacchiotti, y F. Zanzotto, F. 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. En Sirmakessis, S. (ed.) *Knowledge Mining. Studies in Fuzziness and Soft Computing*. Vol. 185, páginas 255-279, Berlín, Springer.

Pazienza, M. 1998: A domain specific terminology extraction system. In: International Journal of Terminology. Benjamin Ed., Vol.5.2 183-201.

Salton, G., y C. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4): 351-372.

Vivaldi, J., y H. Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2): 225-248.

Wüster, E. 1979. *Introduction to the General Theory of Terminology and Terminological Lexicography*. Wien/New York, Springer.