Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 165-168 recibido 24-03-2015 revisado 27-04-2015 aceptado 14-05-2015

Termonet: Construcción de terminologías a partir de WordNet y corpus especializados

Termonet: Terminology construction from WordNet and technical corpora

Miguel Anxo Solla Portela
Universidade de Vigo
Grupo TALG
miguelsolla@uvigo.es

Xavier Gómez Guinovart
Universidade de Vigo
Grupo TALG
xgg@uvigo.es

Resumen: En esta presentación, mostraremos la metodología y los recursos utilizados en el desarrollo de Termonet, una herramienta para la consulta y verificación en corpus de los léxicos de especialidad incluidos en WordNet. Termonet realiza una identificación en WordNet de los synsets pertenecientes a un ámbito terminológico a partir de las relaciones léxico-semánticas establecidas entre los synsets, y valida los términos identificándolos en un corpus especializado desambiguado semánticamente. La construcción de esta herramienta forma parte de las tareas del proyecto de investigación SKATeR-UVigo, orientado al desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego.
Palabras clave: WordNet, lexicografía computacional, terminología computacional

Abstract: In this presentation, we review the methodology and the resources used in the development of Termonet, a tool for checking and verifying in a corpus the specialty lexicons embedded in WordNet. This tool performs an identification of the synsets in WordNet belonging to a terminological domain from the lexical-semantic relations established among synsets, and validates the terms identifying them by means of a semantically disambiguated specialized corpus. The construction of this tool is part of the tasks of the SKATeR-UVigo research project, aimed at the development and application of resources for Galician language processing.
Keywords: WordNet, computational lexicography, computational terminology

1 Introducción

En este artículo[1] se describen la metodología y los recursos utilizados en el desarrollo de Termonet[2], una herramienta para la consulta de los léxicos de especialidad incluidos en WordNet[3] y para su verificación en corpus. La construcción de esta herramienta forma parte de los objetivos del proyecto de investigación SKATeR-UVigo, orientado al desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego.

Termonet se centra en la explotación de WordNet para la construcción de terminologías, mediante la exploración de las relaciones semánticas codificadas entre los nodos conceptuales (o synsets) de la ontología léxica. Como se explica con detalle más adelante, el funcionamiento de la aplicación se basa en que los términos propios de un ámbito terminológico incluidos en WordNet se localizan en synsets relacionados con un nodo raíz mediante ciertas configuraciones de relaciones semánticas y a determinadas distancias máximas de este nodo.

Termonet ofrece la posibilidad de explorar los distintos conjuntos de synsets asociados a un synset de origen en función de las configuraciones definidas por el usuario para la selección de relaciones exploradas y para el nivel máximo de exploración de cada re-

[1]Esta investigación se realiza en el marco del proyecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVigo)* financiado por el Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.
[2]`http://sli.uvigo.es/termonet/termonet.php`
[3]`http://wordnet.princeton.edu`

ISSN 1135-5948

lación. La misma aplicación permite verificar los resultados de la exploración en un corpus de textos especializados.

2 Recursos

Las funcionalidades de Termonet se fundamentan en dos recursos básicos: un léxico WordNet y un corpus textual lematizado y desambiguado con respecto a los sentidos de WordNet. En la implementación actual de Termonet, diseñada para su aplicación en tareas terminológicas relacionadas con la ampliación del WordNet del gallego en el ámbito de la medicina, estos dos recursos son el léxico Galnet y el *Corpus Técnico do Galego*.

Galnet, la versión gallega de WordNet, se distribuye como parte del MCR (González Agirre, Laparra, y Rigau, 2012). Esta versión de Galnet, de 2012, incluye los Basic Level Concepts[4], los ficheros lexicográficos de partes del cuerpo y de substancias, y la traducción parcial de los adjetivos. Además, contiene una primera ampliación realizada con el WN-Toolkit[5] a partir de la Wikipedia[6] y el *Dicionario CLUVI inglés-galego*[7].

A partir de esta versión inicial, se ha seguido ampliando Galnet con el WN-Toolkit a partir de los diccionarios de Apertium[8], Babelnet[9] 2.0, Wiktionary[10], Wikipedia, Geonames[11], Wikispecies[12] y los corpus SemCor inglés-gallego y CLUVI (Gómez Guinovart y O., 2014). También se ha realizado una expansión a partir del *Dicionario de sinónimos do galego*[13] (Gómez Guinovart y Solla Portela, 2014). Finalmente, se han efectuado ampliaciones en el ámbito de la fraseología (locuciones verbales) y de la terminología (medicina y economía). Todas estas expansiones se pueden consultar en la interfaz web de Galnet[14] utilizando la versión de desarrollo del recurso.

La implementación actual de Termonet usa la versión de desarrollo de Galnet 3.0.10 (2015), cuya extensión en número de synsets (Syns) y variantes léxicas (Vars) se recoge en la Tabla 1 en comparación con la de la versión de 2002 distribuida con el MCR.

	MCR		3.0.10	
	Vars	Syns	Vars	Syns
N	18949	14285	27825	20621
V	1416	612	4199	1564
Adj	6773	4415	8086	5104
Adv	0	0	471	370
Total	27138	19312	40581	27659

Tabla 1: Extensión léxica de Galnet

Por su parte, el *Corpus Técnico do Galego* (CTG)[15] es un corpus de orientación terminológica de 15 millones de palabras, formado por textos especializados del gallego contemporáneo en los ámbitos del derecho, informática, economía, ciencias ambientales, ciencias sociales y medicina. La sección del corpus de medicina del CTG (el subcorpus *Medigal*) utilizada en la implementación actual de Termonet totaliza 3.823.232 palabras. Para esta aplicación, se ha utilizado una versión del Medigal etiquetada mediante FreeLing[16] y UKB (Agirre y Soroa, 2009), empleando Galnet 3.0.10 como léxico para la desambiguación semántica del corpus.

3 Funcionalidades

3.1 Construcción de terminologías

La función principal de Termonet consiste en facilitar la extracción de variantes de WordNet relacionadas con un ámbito de especialidad. Con este fin, Termonet ofrece un formulario de consulta que permite elegir un synset de la ontología léxica y, a partir de él, realizar una extracción de los términos relacionados en función de la configuración de relaciones semánticas que se seleccionen. Aunque Termonet permite realizar la extracción desde cualquier synset de la ontología, dada su orientación terminológica, la aplicación trata de sugerir siempre las variantes nominales más próximas cuando se propone un synset no nominal.

Como se ilustra en la parte superior de la Figura 1, Termonet permite indicar el synset de origen que definirá el ámbito de la extracción terminológica, y seleccionar el conjunto de relaciones semánticas que se utilizarán para la identificación de los términos de ese

[4] http://adimen.si.ehu.es/web/BLC/
[5] http://sourceforge.net/projects/wn-toolkit/
[6] http://www.wikipedia.org
[7] http://sli.uvigo.es/dicionario/
[8] http://www.apertium.org
[9] http://www.babelnet.org
[10] http://www.wiktionary.org
[11] http://www.geonames.org
[12] http://species.wikimedia.org
[13] http://sli.uvigo.es/sinonimos/
[14] http://sli.uvigo.es/galnet/
[15] http://sli.uvigo.es/CTG/
[16] http://nlp.lsi.upc.edu/freeling/

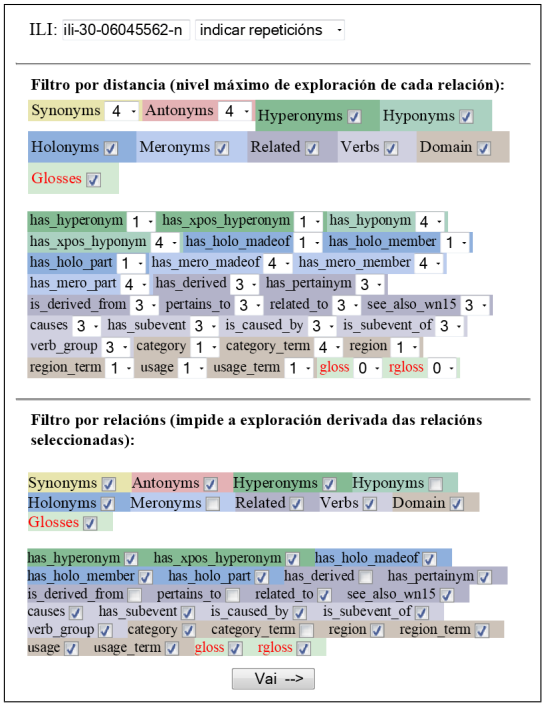

Figura 1: Consulta en Termonet.

Figura 2: Extracción de terminología.

ámbito, así como la distancia o nivel de profundidad hasta donde se desea desplegar cada tipo de relación. El concepto de distancia se refiere aquí al número de relaciones léxico-semánticas que unen dos synsets entre sí. De este modo, Termonet desplegará el árbol de relaciones desde el synset de origen a través de esa relación hasta alcanzar el nivel de profundidad determinado. Véase en la Figura 2, por ejemplo, la relación de hiponimia desplegada hasta el nivel 4 de profundidad en la terminología del ámbito de la medicina, construida a partir del synset *medical science* con los parámetros ilustrados en la Figura 1.

La aplicación cuenta también con un subformulario (parte inferior de la Figura 1) que permite restringir la extracción terminológica impidiendo la exploración derivada de las relaciones semánticas seleccionadas. Mediante este filtro, se trata de limitar la *toxicidad* de ciertas relaciónes semánticas para la selección de los términos de un ámbito de especialidad, es decir, de reducir el impacto de las relaciones que introducen synsets que se desvían del campo conceptual. Según este criterio, la hiperonimia, por ejemplo, se suele considerar una relación *tóxica*, ya que amplía la cobertura semántica inicial y tiende a introducir términos de campos conceptuales más amplios que los de partida.

Aunque la herramienta de extracción terminológica se encuentra aún en fase de desarrollo, en los experimentos se obtuvieron, con configuraciones muy simples de los parámetros de extracción, conjuntos de resultados con una congruencia mayor y cuantitativamente más significativos que la selección de variantes ligadas a un dominio de WordNet Domains[17]. Además, la extracción puede partir de cualquier synset y no está limitada a un dominio preestablecido, de modo que el procedimiento es idéntico para ámbitos conceptuales amplios, como la biología, y para campos más concisos, como la microbiología.

3.2 Verificación en corpus

Como ya se ha mencionado anteriormente, Termonet permite verificar los resultados de la extracción en un corpus textual lematizado y desambiguado con respecto a los sentidos de WordNet. En su implementación actual, permite contrastar los términos gallegos identificados en el corpus de medicina Medigal etiquetado con FreeLing y UKB.

El corpus desambiguado facilita el desarollo de estrategias de verificación con base semántica para las variantes monoléxicas procedentes de Galnet, pero no para las pluriléxicas, que no cuentan con etiquetación semánti-

[17] `http://wndomains.fbk.eu`

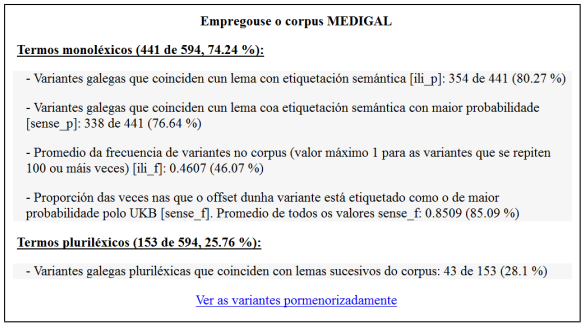

Figura 3: Verificación en corpus.

ca debido a las características de la lematización del corpus con FreeLing. Con el fin de comprobar de algún modo su presencia en el corpus, Termonet identifica las palabras léxicas de la variante en lemas sucesivos del corpus y calcula su frecuencia.

Termonet evalúa la presencia de cada término monoléxico en el corpus en base a cuatro criterios cuantificados de 0 a 1, y finalmente combina los resultados obtenidos por todos ellos en un índice general para cada criterio. Los criterios aplicados son:

1. La variante está presente (1) o no (0) como lema del corpus y con la etiqueta semántica del synset correspondiente.
2. La variante está presente como lema del corpus y con la etiqueta semántica más probable (1) o no (0) según UKB.
3. Frecuencia absoluta de la variante en el corpus, ponderando el valor máximo (1) para las variantes etiquetadas semánticamente que se repiten 100 veces o más, y el valor mínimo (0) para las variantes que no están presentes en el corpus.
4. Frecuencia con la que UKB le atribuye la mayor probabilidad a la etiqueta del synset de la variante, asignando el valor máximo (1) para la totalidad de las veces y el mínimo (0) para ninguna.

En la Figura 3 se muestran los índices globales obtenidos por la terminología construida a partir del synset *medical science* con los parámetros ilustrados en la Figura 1. A partir del análisis pormenorizado de las variantes (Figura 4), Termonet ofrece la posibilidad de comprobar sus contextos de uso en el corpus especializado (Figura 5), permitiendo así adquirir información terminológica muy valiosa sobre el uso real de los términos.

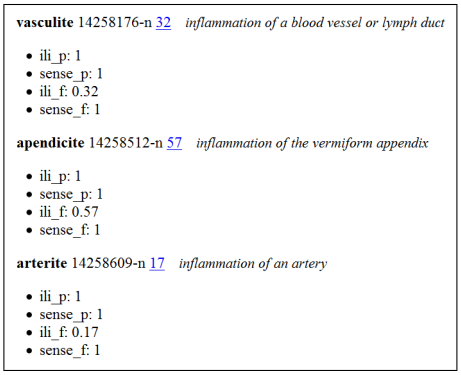

Figura 4: Evaluación de los términos.

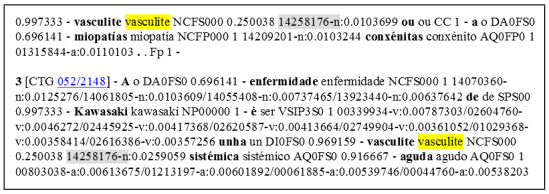

Figura 5: Término en contexto.

4 *Conclusiones*

La verificación de los términos en un corpus desambiguado permite adquirir información muy valiosa sobre su uso real y constituye una fuente de conocimiento muy relevante en la expansión de Galnet guiada por campos conceptuales. Los resultados obtenidos en la extracción, avalados por su evaluación en corpus, nos animan a continuar investigando en esta dirección y a seguir completando el WordNet del gallego desde esta perspectiva.

Bibliografía

Agirre, E. y A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the ACL*, págs. 33–41.

Gómez Guinovart, X. y Antoni O. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50.

Gómez Guinovart, X. y M. A. Solla Portela. 2014. O dicionario de sinónimos como recurso para a expansión de WordNet. *Linguamática*, 6(2):69–74.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *6th Global WordNet Conference*.