Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos

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

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

Resumen

En este artículo mostramos las técnicas utilizadas y los primeros resultados de un experimento de expansión del WordNet gallego mediante extracción léxica a partir de un diccionario de sinónimos de esta lengua.

1. Introducción

El objetivo del experimento presentado en este trabajo[1] es la expansión del WordNet gallego mediante extracción léxica a partir de un diccionario de sinónimos de esta lengua. El experimento se realiza dentro del proyecto coordinado SKATeR en el que nuestro grupo tiene como objetivo prioritario la construcción de Galnet, la versión gallega del WordNet 3.0. El marco de desarrollo en el que se integra Galnet es el Multilingual Central Repository (MCR) [4], una plataforma que abarca los léxicos WordNet de cinco lenguas (inglés, español, catalán, vasco y gallego) enlazados por el índice interlingüístico (ILI) correspondiente al WordNet 3.0 y con los synsets categorizados en la jerarquía de dominios IRST y en las ontologías SUMO y Top Concept Ontology.

Galnet se distribuye con licencia Creative Commons como parte del MCR[2]. La versión de Galnet de esta distribución alcanza la cobertura léxica que se muestra en la Tabla 1 en comparación con la del WordNet 3.0 del inglés.

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	18949	14285
V	25047	13767	1416	612
Adj	30002	18156	6773	4415
Adv	5580	3621	0	0
TOTAL	206941	117659	27138	19312

Cuadro 1: Distribución actual de Galnet en el MCR

Esta primera distribución pública de Galnet (de finales del 2012) se inició con la traducción al gallego de los synsets nominales y verbales pertenecientes a los Basic Level Concepts (BLC). Más concretamente, se tradujeron y

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

[1]Esta investigación se ha llevado a cabo gracias al 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://adimen.si.ehu.es/web/MCR/`

adaptaron al gallego los 649 synsets nominales y 616 synsets verbales agrupados en freqmin20/all en la distribución oficial de los BLC para WordNet 3.0[3]. Esta versión inicial de Galnet incluye también la traducción gallega de los ficheros lexicográficos correspondientes a las partes del cuerpo (noun.body), a las substancias (noun.substance) y a los adjetivos de tipo general (adj.all)[4]. Así mismo, esta versión incluye una primera expansión realizada con el WN-Toolkit [5] que amplía la cobertura léxica de Galnet a partir de dos recursos bilingües inglés-gallego ya existentes, la Wikipedia y el Dicionario CLUVI Inglés-Galego[5].

A partir de esta versión base de 2012, se ha seguido ampliando Galnet mediante técnicas de extracción léxica basadas en recursos textuales bilingües existentes. Concretamente, se ha llevado a cabo una nueva extracción léxica con WN-Toolkit a partir de los corpus paralelos CLUVI[6] y SemCor[7] y de diversos léxicos bilingües (Apertium[8], Wiktionary[9] y Babelnet[10][3]. Igualmente, se ha empezado a trabajar en la extracción léxica a partir del *Dicionario de sinónimos do galego*[11], único diccionario electrónico del gallego de este tipo, con una extensión actual de 27.104 entradas, 44.849 acepciones y 203.251 sinónimos[2].

Los resultados de estas últimas expansiones en curso, aún en fase de completar la revisión e introducción de la extracción, se pueden ver en la interface de consulta de Galnet[12] realizando las consultas sobre la versión de desarrollo del recurso, cuya cobertura actual se muestra en en la Tabla 2

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	20740	15581
V	25047	13767	3568	1239
Adj	30002	18156	7627	4809
Adv	5580	3621	167	153
TOTAL	206941	117659	32102	21782

Cuadro 2: Cobertura actual de Galnet (versión de desarrollo 3.0.2)

Aunque ya se realizó previamente un experimento de extracción a partir del *Dicionario de sinónimos do galego*[1], el objetivo de insistir en la experimentación con el mismo recurso no es otro que tratar de obtener una mayor cobertura de variantes para el Galnet a través de la extracción automática de candidaturas procedentes del diccionario de sinónimos, que cuenta con un volumen considerable de lemas en su repertorio organizados semánticamente con ciertas similitudes respecto al Galnet. Los resultados de los experimentos anteriores no cerraron las puertas para insistir en el reaprovechamiento de la obra lexicográfica, sino que dejaron entrever que tal vez se podría plantear una nueva hipótesis con el fin de rentabilizar más eficazmente una extracción para alimentar la nueva versión del WordNet en lengua gallega, tanto cuantitativa como cualitativamente.

2. Experimento

2.1. Método propuesto

El análisis de resultados de un experimento anterior para la incorporación de lemas del *Dicionario de sinónimos do galego* revelaba que todavía se podrían intentar nuevas estrategias para explotar el caudal léxico y la organización semántica del diccionario. Este experimento previo se diseñó con el objetivo de identificar los lemas del diccionario y las variantes de Galnet con un bajo índice de frecuencia que fuesen idénticos. La hipótesis de partida consistía en que los lemas que aparecen en muy pocas ocasiones tienen mayor probabilidad de identificar formas monosémicas y, por lo tanto, al encontrarse tanto en el diccionario como en Galnet permitiría trasladar, tras una revisión humana, los sinónimos correspondientes a esa acepción lexicográfica como variantes del mismo synset en el WordNet gallego. De entre las variantes que se documentaron una única vez en las dos obras (hápax

[3] http://adimen.si.ehu.es/web/BLC/
[4] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html
[5] http://sli.uvigo.es/dicionario/
[6] http://sli.uvigo.es/CLUVI/
[7] http://www.gabormelli.com/RKB/SemCor_Corpus/
[8] http://sourceforge.net/projects/apertium/
[9] http://www.wiktionary.org
[10] http://babelnet.org
[11] http://sli.uvigo.es/sinonimos
[12] http://sli.uvigo.es/galnet/

legómena), una vez que se realizó la revisión lexicológica, se aprobaron el 65 % de las 4.283 candidaturas producto del cruce automático[1].

En el momento en que se finalizó este experimento y se importaron las nuevas variantes para Galnet el diccionario se había revisado y en el WordNet del gallego ya se habían importado los resultados de otros experimentos, por lo que se abría la posibilidad de repetir el experimento tratando de mejorar la eficacia y reducir la intervención humana. Con el fin disminuir el coste de la revisión lexicológica y de extraer aun más información de forma automática se diseñó una nueva estrategia que se basa en la hipótesis de que si al menos dos lemas son sinonímicos en la misma acepción de la obra lexicográfica y eses mismos lemas son variantes del mismo synset en WordNet, es probable que se trate del mismo sentido; es decir, que la acepción lexicográfica refleje el mismo valor semántico que el synset de WordNet. En consecuencia, las formas sinonímicas restantes de la acepción lexicográfica son susceptibles de convertirse en variantes del mismo synset y ampliar la cantidad de variantes presentes en el Galnet en estos casos.

La organización interna del diccionario de sinónimos utilizado para el experimento parte de un total de 203.251 lemas (que no suelen ser únicos, sino que a menudo se repiten en diferentes entradas y/o acepciones). La versión en desarrollo de Galnet cuenta en la actualidad con 32.102 variantes (que también se pueden repetir en diferentes synsets, aunque con un índice de frecuencia en la repetición sensiblemente menor que en el caso del diccionario).

2.2. Resultados

Tras el cruce automático de dos sinónimos en la misma acepción del diccionario con la misma categoría gramatical y con dos variantes idénticas en el mismo synset tanto en el diccionario como en Galnet, se han obtenido 25.186 candidaturas diferentes a constituir variantes nuevas, cada una de ellas asociada al synset correspondiente, para enriquecer el WordNet del gallego. Pese al optimismo que produce la obtención de un alto número de candidaturas, su introducción en la red léxico-semántica que conforma WordNet necesita de una revisión lexicológica que garantice la congruencia de cada synset. Ante una cantidad tan ingente de propuestas para revisar, se intentó obtener una verificación de la hipótesis de partida utilizando la misma metodología con una fuente distinta; así mismo, se diseñó una repetición del experimento en fases con el objetivo de limitar la cantidad de resultados y mejorar la precisión de las propuestas.

Con el fin de verificar la validez del método propuesto, se realizó la misma prueba con una fuente diferente, un thesaurus elaborado a partir de las sinonimias que ofrece el Vocabulario Ortográfico da Lingua Galega (VOLGa)[13]. Las características de este thesaurus son muy diferentes a las del diccionario, pues cuenta unicamente con 6.960 sinónimos organizados en 3.263 synsets, motivo por el que la cantidad de synsets con más de dos formas sinónimicas no es muy numerosa y la probabilidad de que los sinónimos no se encuentren ya en Galnet es también reducida; sin embargo, al tratarse de una fuente normativa, se incrementa su fiabilidad. Como producto de esta última prueba se obtuvieron solamente 42 candidaturas a variantes nuevas para Galnet y tras su revisión lexicológica únicamente 4 formas candidatas fueron rechazadas por una asignación incorrecta del valor semántico del synset que tenían asignado.

Para restringir la cantidad de candidaturas procedentes del diccionario de sinónimos se rediseñó el experimento dividiéndolo en fases que permitiesen la revisión humana en plazos de tiempo más razonables. Se repitió el experimento cruzando las acepciones del diccionario que compartiesen 3 sinónimos o más con 3 variantes en el mismo synset de Galnet y se obtuvieron 6.335 candidaturas. Para evaluar la adaptación de los resultados en Galnet se efectuó una cata de las últimas 100 formas candidatas a variantes y se realizó una revisión lexicológica de cada una de ellas.

Tras esta revisión se confirmó que la precisión de las formas candidatas obtenidas automáticamente era relativa, pues sólo el 35 % de las candidaturas se consideraron correctas a causa de diferentes factores: por una parte, factores formales derivados de las características del diccionario de sinónimos, pues esta obra lexicográfica, ideada originariamente para el sector editorial, contiene formas dialectales, variantes que no son normativas, popularismos, formas con interferencias lingüísticas, etc.; por otra parte, factores debidos a la mala asignación conceptual en casos de polisemia.

Así mismo, durante la revisión de las formas candidatas, se detectó que la precisión disminuía según se incrementaba el índice de dispersión semántica; es decir, que cuando existe un número de elevado de sinónimos en la misma acepción del diccionario, las candidaturas propuestas para incorporarse a Galnet son menos acertadas. Como fruto de esta observación se repitió el experimento con el cruce de tres formas sinonímicas que coincidan con tres variantes con la misma categoría gramatical entre sí y que además se limitase a las acepciones del

[13] `http://www.realacademiagalega.org/recursos-volg/`

diccionario que no tuviesen más de 5 sinónimos. El resultado fue de 856 formas candidatas a variantes de las que se seleccionó una cata con las 100 primeras para su revisión. El índice de precisión de esta cata es ligeramente superior al 60 % y constituye un punto de partida asumible para una revisión humana eficaz. Dado que la metodología que se ha utilizado admite sin lugar a dudas la recursividad (tras cada ampliación de Galnet el cruce de sinónimos y variantes puede ofrecer nuevos resultados presumiblemente más precisos), el experimento se irá repitiendo en fases sucesivas que vayan ampliando la cobertura de los cruces, durante las cuales se irán eliminando paulatinamente las restricciones que se han descrito, y se establecerá un nuevo filtro para que no se generen candidaturas idénticas a las que no hayan sido aceptadas en revisiones humanas anteriores.

3. Conclusiones

Un mero análisis cuantitativo de los resultados podría reflejar la posibilidad de un gran aumento en el WordNet gallego si se corrobora la incorporación de la mayor parte de las candidaturas a variantes procedentes de la extracción del diccionario de sinónimos, sin embargo todas estas candidaturas enriquecen synsets que ya tenían al menos dos variantes previas en el Galnet y en contadas ocasiones amplian la cobertura (únicamente en algunos casos debido a la intervención humana durante la revisión) a nuevos synsets o a synsets que tienen una única variante. Por lo tanto, es necesario relativizar el impacto que pueda suponer la inclusión de estas variantes nuevas, pues uno de los objetivos principales del grupo de investigación es ampliar Galnet en todas las dimensiones y es preciso considerar que es complementario de otros experimentos que inciden en la ampliación de WordNet ofreciendo variantes para los synsets en los que todavía no se ha introducido ninguna.

Cabe destacar también que en el momento en que se redacta esta comunicación los resultados están pendientes todavía de una revisión más amplia desde una perspectiva lexicológica y que el desarrollo del experimento se encuentra en fase inicial. Además, la evolución de la experimentación podría indicar posibles mejoras en el índice de precisión, pues el factor humano durante la revisión lexicográfica de los resultados tiene un peso determinante en la metodología dada la gran cantidad de candidaturas.

Para concluir, pensamos que esta metodología de expansión de WordNet podría aplicarse sin demasiadas modificaciones en proyectos de ampliación de WordNet en otros idiomas, siempre que se disponga para la lengua de repertorios léxicos con características similares al *Dicionario de sinónimos do galego* utilizado para esta investigación.

Referencias

[1] Gómez Guinovart, Xavier: Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego. En Ana Gabriela Macedo, Carlos Mendes de Sousa, Vítor Moura (eds.), XV Colóquio de Outono - As humanidades e as ciências: disjunções e confluências. CEHUM: Universidade do Minho. (2014)

[2] Gómez Guinovart, Xavier y Alberto Simões: Retreading Dictionaries for the 21st Century. En José Paulo Leal, Ricardo Rocha y Alberto Simões (eds.), 2nd Symposium on Languages, Applications and Technologies. OASIcs: Open Access Series in Informatics, vol. 29. Dagstuhl Publishing: Saarbrücken. (2013) 115-126.

[3] Gómez Guinovart, Xavier y Antoni Oliver: Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. XXX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. Girona. (2014)

[4] González Agirre, Aitor y German Rigau: Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. Linguamática, 5.1. (2013) 13-28.

[5] Oliver, Antoni: WN-Toolkit: Automatic generation of WordNets following the expand model. Proceedings of the 7th Global WordNet Conference. Tartu, Estonia. (2014)