O dicionario de sinónimos como recurso para a expansión de WordNet*

The dictionary of synonyms as a resource for expanding WordNet

Xavier Gómez Guinovart
Universidade de Vigo
xgg@uvigo.es

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

Resumo

Neste artigo presentamos os alicerces dun experimento de extracción léxica deseñado no marco do proxecto de investigación SKATeR e orientado á ampliación do WordNet do galego mediante a explotación dos datos lexicográficos recollidos nun dicionario de sinónimos "tradicional" desta lingua.

Palabras chave

WordNet, adquisición de información léxica, dicionario de sinónimos, recursos lingüísticos

Abstract

In this paper, we present the foundations for a lexical acquisition experiment designed in the framework of the SKATeR research project and aimed to the expansion of the Galician WordNet using the lexicographical data collected in a "traditional" Galician dictionary of synonyms.

Keywords

WordNet, lexical acquisition, dictionary of synonyms, language resources

1 Introdución

Neste artigo[1] presentamos os alicerces dun experimento de extracción léxica deseñado no marco do proxecto de investigación SKATeR[2] e orientado á ampliación do WordNet do galego mediante a explotación dos datos lexicográficos recollidos nun dicionario de sinónimos "tradicional" desta lingua.

En principio, as características de WordNet (Miller et al., 1990) como rede léxico-semántica estruturada en nós conceptuais (*synsets*) compostos por *variantes* léxicas dun mesmo significado permitirían inferir sen dificultades a hipótese de que os dicionarios de sinónimos existentes dunha lingua constitúen fontes lexicográficas directas moi axeitadas para a expansión deste recurso.

Porén, os experimentos previos de extensión do WordNet do galego a partir dun dicionario de sinónimos realizados polos autores (Gómez Guinovart, 2014) (Solla Portela e Gómez Guinovart, 2014) demostran as dificultades de acadar mediante este recurso uns índices de precisión que permitan a validacion manual dos resultados da extracción nun tempo razoábel. As causas desta dificultade radican principalmente no distinto concepto de sinonimia utilizado por cada un destes dous recursos, moito máis estrito e delimitado pola glosa no WordNet (Gómez Clemente et al., 2013), moito máis laxo e abranguente doutras relacións semánticas no dicionario de sinónimos tradicional.

Deste xeito, mentres que o concepto de sinonimia manexado en WordNet fai referencia ao concepto máis específico de sinonimia contextual ou intercambiabilidade dos sinónimos limitada a un contexto (Miller, 1998), no dicionario de sinónimos tradicional as relacións semánticas entre o lema e os "sinónimos" que forman parte da entrada pode ser tanto de sinonimia, como de hiperonimia, hiponimia, holonimia, meronimia, troponimia ou implicación, entre outras.

Así, por exemplo, no dicionario de sinónimos usado neste experimento, a entrada para o lema *climaterio* ("período da vida do home e da muller en que se producen unha serie de cambios no organismo debidos á diminución da actividade das glándulas sexuais") está formada polos seus dous hipónimos *andropausa* (isto é, o climaterio mas-

*Esta investigación realízase no marco do proxecto *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 polo Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.

[1]Queremos agradecer aquí sinceramente as valiosas contribucións para a mellora do artigo das tres persoas expertas –Hugo Gonçalo Oliveira, da Universidade de Coimbra; Álvaro Iriarte Sanromán, da Universidade do Minho; e Mercè Lorente Casafont, da Universitat Pompeu Fabra– que realizaron a revisión previa á súa aceptación por parte da revista.

[2]http://nlp.lsi.upc.edu/skater/

culino) e *menopausa* (ou climaterio feminino).

Por este motivo, estamos a traballar no desenvolvemento dunha metodoloxía de extracción que nos permita maximizar os esforzos dedicados á revisión humana dos resultados, aumentando a precisión dos resultados e controlando ao mesmo tempo a súa amplitude, e combinando aspectos tratados nos experimentos previos, como a categoría gramatical e a dispersión semántica das entradas lexicográficas, con outros aínda non tratados, como a frecuencia nas obras lexicográficas dos lemas procesados.

Un traballo semellante de enriquecemento dunha ontoloxía léxica semellante a WordNet a partir doutros recursos léxicos é o realizado para o portugués no proxecto Onto.PT[3] (Gonçalo Oliveira e Gomes, 2014), no que se obtiveron synsets a partir de dicionarios de sinónimos e outras fontes textuais (corpus, enciclopedias e dicionarios da lingua) usando como método de extracción diversos índices de semellanza (Gonçalo Oliveira, 2013).

Nos seguintes apartados, presentaremos os recursos procesados no experimento, a metodoloxía deseñada para a extracción e unha avaliación dos seus resultados.

2 Recursos

Galnet, a versión galega de WordNet, distribúese con licenza Creative Commons como parte do MCR[4] (González-Agirre e Rigau, 2013) (González-Agirre, Laparra e Rigau, 2012). A versión de Galnet desta distribución (de finais de 2012) alcanza a cobertura léxica que se amosa na Táboa 1 (onde a columna *Vars* indica o número total de *variantes* sinonímicas en cada categoría e a columna *Syns* o número total de *synsets* ou nós conceptuais recompilados) en comparación coa do WordNet 3.0 do inglés (na táboa, *EWN30*).

	EWN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	18949	14285
V	25047	13767	1416	612
Adx	30002	18156	6773	4415
Adv	5580	3621	0	0
TOTAL	206941	117659	27138	19312

Táboa 1: Distribución actual de Galnet no MCR

A partir desta versión inicial de 2012, continuamos ampliando Galnet mediante técnicas de extracción léxica baseadas en recursos textuais monolingües e bilingües existentes (corpus paralelos e dicionarios) (Gómez Guinovart e Simões, 2013) (Gómez Guinovart, 2014) (Solla Portela e Gómez Guinovart, 2014) (Gómez Guinovart e Oliver, 2014). Os resultados das expansións en curso pódense observar na interface web de consulta de Galnet[5] realizando buscas sobre a versión de desenvolvemento do recurso. O experimento aquí presentado realizouse coa versión de desenvolvemento 3.0.4 de Galnet, cuxa cobertura se recolle na Táboa 2.

	WN30		Galnet	
	Vars	Syns	Vars	Syns
N	146312	82115	22186	16812
V	25047	13767	3996	1423
Adx	30002	18156	7884	4962
Adv	5580	3621	253	223
TOTAL	206941	117659	34319	23420

Táboa 2: Versión de desenvolvemento 3.0.4

O *Dicionario de sinónimos do galego* (Gómez Clemente, Gómez Guinovart e Simões, 2014) usado para a expansión de Galnet foi elaborado tamén no marco deste proxecto (Gómez Guinovart, 2014) (Gómez Guinovart e Simões, 2013). Este dicionario de sinónimos está dispoñíbel para a súa libre consulta na web[6], pódese descargar tamén como app para Android[7] e para iOS[8], e constitúe o único recurso electrónico existente para a lingua galega dentro desta categoría de dicionarios. No experimento presentado utilizouse a versión de desenvolvemento 0.7 deste recurso coa cobertura léxica que se indica na Táboa 3.

Entradas	27104
Acepcións	44849
Sinónimos	203251

Táboa 3: Extensión do dicionario

Nesta descrición dos contidos do dicionario de sinónimos, entendemos por *entradas* os artigos lexicográficos (na tradición escrita, ordenados alfabeticamente) nos que se divide a estrutura principal do dicionario; entendemos por *acepcións* os conxuntos dun ou máis sinónimos, agrupados por significado, nos que se dividen as entradas; e entendemos por *sinónimos* os elementos léxicos

[3] `http://ontopt.dei.uc.pt/`

[4] `http://adimen.si.ehu.es/web/MCR/`

[5] `http://sli.uvigo.es/galnet/`

[6] `http://sli.uvigo.es/sinonimos/`

[7] `https://play.google.com/store/apps/details?id=net.ayco.sinonimosgal`

[8] `https://itunes.apple.com/us/app/sinonimos-do-galego/id940045971?l=es&ls=1&mt=8`

(monoléxicos ou pluriléxicos) que compoñen as acepcións nas que se dividen as entradas.

3 Metodoloxía

3.1 Alicerces

Temos, por unha banda, un dicionario de sinónimos D formado polo conxunto de acepcións ($\{A_1, A_2, ... A_n\}$) que forman parte da microestrutura das entradas do dicionario

$$D = \{A_1, A_2, ..., A_n\}$$

e temos, por outra banda, un WordNet do galego G formado por un conxunto de synsets ($\{S_1, S_2, ... S_n\}$) que recollen os sinónimos (ou variantes sinónimicas) asignados a cada concepto desta rede léxico-semántica

$$G = \{S_1, S_2, ..., S_n\}$$

Cada acepción A_k de D está formada por un conxunto de formas léxicas sinonímicas ($\{s_1, s_2, ... s_n\}$) composto polo lema e os sinónimos dunha acepción

$$A_k = \{s_1, s_2, ..., s_n\}$$

e cada synset S_l de G está formado por un conxunto de variantes sinonímicas ($\{v_1, v_2, ... v_n\}$) para un concepto da rede léxico-semántica

$$S_l = \{v_1, v_2, ..., v_n\}$$

O método de extracción utilizado baséase na hipótese de que unha acepción lexicográfica A_k do dicionario representa probabelmente o mesmo valor semántico que un synset S_l de Galnet se A_k e S_l comparten cando menos un elemento idéntico da mesma categoría morfolóxica

$$\{s_x, v_y\} \in A_k \cap S_l \wedge s_x = v_y \wedge CAT(s_x) = CAT(v_y)$$

Se se cumpre esta condición, os sinónimos de A_k distintos de s_x (e ausentes de S_l) serán candidatos susceptíbeis de engadirse a S_l após un certo grao de revisión humana, incrementando deste modo o número de variantes do synset.

Considérese, por exemplo, a entrada para o adxectivo *aleuto* no dicionario, presentada no seu código fonte en XML na Listaxe 1, que está conformada por unha única acepción e catro sinónimos. Esta acepción estaría composta por cinco formas léxicas sinonímicas: {*aleuto, agudo, espelido, intelixente, listo*}.

Por outra banda, considérese o synset adxectivo identificado como glg-30-00061885-a formado no Galnet 3.0.4 polo conxunto de dúas variantes sinonímicas {*enxeñoso, aleuto*}.

De acordo coa metodoloxía utilizada, supoñemos que esta acepción do dicionario de sinónimos para a entrada *aleuto* e este synset glg-30-00061885-a do Galnet son susceptíbeis de representar o mesmo concepto semántico, xa que as dúas constelacións léxicas son de categoría morfolóxica adxectiva e as dúas comparten unha forma coincidente (*aleuto*). Deste xeito, os sinónimos da entrada *aleuto* do dicionario ausentes do synset glg-30-00061885-a do Galnet –isto é, os sinónimos {*agudo, espelido, intelixente, listo*}– serán variantes candidatas a integrar o synset glg-30-00061885-a do Galnet previa revisión humana.

Listaxe 1: Entrada de *aleuto*

```
<entry>
<form>
<orth>aleuto</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1">
<syn><lemma>Agudo</lemma></syn>,
<syn><lemma>espelido</lemma></syn>,
<syn><lemma>intelixente</lemma></syn>,
<syn><lemma>listo</lemma></syn>.
</def>
</sense>
</entry>
```

3.2 Parámetros

O problema desta condición mínima é que produce moito ruído: ofrece moitos sinónimos candidatos para formar parte dun synset (alta cobertura) pero o seu índice de acertos non é moi elevado (baixa precisión), o que imposibilita a planificación dunha revisión humana dos resultados. Concretamente, cos recursos anteriormente descritos, o número de candidaturas elévase a 296.246 sinónimos.

Por esta razón, deseñamos un experimento que nos permitise maximizar os esforzos (sempre demasiado escasos) dedicados á revisión humana dos resultados, aumentando o máis posíbel a precisión dos resultados sen diminuír a súa cobertura a límites inútiles. Os parámetros que se tiveron en conta neste experimento foron os seis que se indican deseguido:

- P_1 Número de elementos idénticos e coa mesma categoría morfolóxica en $A_k \cap S_l$ (mínimo 1)
- P_2 Número de elementos en A_k
- P_3 Frecuencia absoluta do sinónimo candidato en D

P_4 Frecuencia absoluta do sinónimo candidato en G

P_5 Frecuencia absoluta en D das formas léxicas compartidas (idénticas e coa mesma categoría morfolóxica) en A_k e S_l

P_6 Frecuencia absoluta en G das formas léxicas compartidas (idénticas e coa mesma categoría morfolóxica) en A_k e S_l

P_1 determina a cantidade de formas compartidas entre o synset e a acepción. Se A_k e S_l comparten algún elemento coa mesma forma e coa mesma categoría gramatical existe a posibilidade de que compartan o significado. En principio, cantos máis elementos compartan A_k e S_l, maior será a seguridade de atinar na súa coincidencia semántica e, por tanto, a precisión dos resultados será maior. Por outra banda, a cobertura do experimento descende a medida que sobe P_1. Na Táboa 4 amósase o número de candidatos resultantes ao usarmos só esta condición.

P_1	candidatos
> 0	296.246
> 1	26.610
> 2	6.698
> 3	2.954
> 4	1.421
> 5	786
> 6	436
> 7	220
> 8	178
> 9	135

Táboa 4: Candidatos por P_1

Con P_2 tratamos de manexar o fenómeno da dispersión semántica propio do dicionario de sinónimos. O certo é que, durante a revisión das formas candidatas que se obtiveron nun experimento anterior (Solla Portela e Gómez Guinovart, 2014), detectouse que a precisión das candidaturas propostas para se incorporaren a Galnet diminuía a medida que se incrementaba o número de sinónimos na mesma acepción do dicionario, debido probabelmente á menor cohesión semántica entre as formas agrupadas na acepción. De xeito experimental, puidemos comprobar que os valores de P_2 entre 6 e 9 son os que ofrecen mellores resultados para incrementar a precisión sen que baixe demasiado a cobertura. Tomando como referencia un valor maior ca 1 para P_1, recollemos na Táboa 5 o número de candidatos que se obtiveron como resultado con diferentes valores de P_2.

P_{3-6} permítennos tratar no experimento o fenómeno da polisemia das formas léxicas que se manexaron. A idea é que cantas máis veces apareza unha mesma forma léxica no recurso léxico analizado (D ou G), maior será a posibilidade de que se trate dunha forma léxica polisémica e, polo tanto, maior será a posibilidade de incidir negativamente na selección das candidaturas. Por unha banda, P_{3-4} permiten controlar a frecuencia das formas candidatas en D e en G; e, por outra banda, P_{5-6} permiten controlar a frecuencia das formas compartidas entre D e G durante a selección das candidatas. En ambos os casos, o factor de control que ofrece maior rendemento é o que se aplicou sobre D (isto é, P_3 e P_5) e, asemade, a supervisión da frecuencia das formas candidatas (P_3 e P_4) móstrase máis eficaz que a das formas compartidas (P_5 e P_6). Por exemplo, tomando como punto de partida a táboa anterior, con $P_1 > 1$ e con $P_2 < 8$, a Táboa 6 recolle o número de candidaturas que se obtiveron cos diferentes valores usados no experimento para P_3.

P_2	candidatos
< 3	0
< 4	1.088
< 5	3.895
< 6	7.265
< 7	10.771
< 8	14.010
< 9	16.317
< 10	18.347
< 11	19.815
< 12	20.995
< 13	21.861

Táboa 5: Candidatos por P_2 para $P_1 > 1$

P_3	candidatos
< 12	7.262
< 11	6.716
< 10	6.141
< 9	5.588
< 8	4.944
< 7	4.292
< 6	3.585
< 5	2.870
< 4	2.070
< 3	1.292
< 2	474

Táboa 6: Candidatos por P_3 para $P_1 > 1$ e $P_2 < 8$

4 Avaliación

A avaliación da precisión dos resultados realizouse mediante a validación manual dos candidatos. Realizamos diversos experimentos de extrac-

experimento	P_1	P_2	P_3	P_4	P_5	P_6	candidatos	precisión
E_1	> 1	< 9	< 3	-	< 4	-	60	47 %
E_2	> 1	< 9	< 8	< 2	< 5	< 2	95	22 %
E_3	> 1	< 9	< 5	< 12	< 4	< 12	85	19 %
E_4	> 1	< 4	< 13	< 13	< 4	< 13	29	14 %
E_5	> 1	< 4	< 4	< 13	< 13	< 13	77	39 %
E_6	> 1	< 4	< 8	< 8	< 8	< 8	92	37 %
E_7	> 2	< 9	< 8	< 12	< 8	< 12	52	20 %
E_8	> 2	-	-	-	-	-	6.335	35 %
E_9	> 2	< 6	-	-	-	-	856	60 %

Táboa 7: Precisión

ción usando diferentes combinacións de parámetros, tratando sempre de acadar un compromiso entre cobertura e precisión que nos permita maximizar a rendibilidade das horas de dedicación humana á tarefa de revisión. A Táboa 7 presenta os datos de precisión en diversas combinacións de parámetros que se experimentaron cos seguintes criterios:

E_1 Dispersión moi baixa e frecuencia baixa no dicionario de sinónimos das formas compartidas, e sen restricións na frecuencia en Galnet

E_2 Frecuencia mínima en Galnet das formas candidatas e compartidas

E_3 Frecuencia baixa no dicionario das formas candidatas e compartidas

E_4 Dispersión baixa e frecuencia baixa no dicionario das formas compartidas

E_5 Dispersión baixa e frecuencia baixa no dicionario das formas candidatas

E_6 Baixa dispersión semántica

E_7 Coincidencia de 3 formas no dicionario e en Galnet

Para a súa comparación, inclúense os resultados previos de $[E_8]$ e $[E_9]$ que se obtiveron coa versión 3.0.2 de Galnet a partir da avaliación manual de 100 formas candidatas (Solla Portela e Gómez Guinovart, 2014).

Os datos que se analizan na Táboa 7 reflicten o impacto inicial da aplicación da metodoloxía e da comparación entre as parametrizacións expostas nun estadio concreto do desenvolvemento do proxecto; mais prevese unha mellora da precisión en futuras reutilizacións do procedemento baseándose en (a) a escolla de parámetros similares aos que ofreceron maior rendibilidade neste experimento, (b) a incorporación dun filtro de candidaturas non-desexadas a partir das formas desbotadas na revisión humana e (c) o cruzamento dos sinónimos cunha versión aumentada e revisada do Galnet respecto da que se utilizou nesta ocasión.

5 Conclusións

Os resultados deste experimento destacan a necesidade de lograr un compromiso entre cobertura e precisión que facilite a viabilidade da revisión humana.

A pesar de que a precisión que se obtivo foi máis ben baixa na maior parte dos experimentos con diversas combinacións de parámetros, prevese que a aplicación cíclica dos experimentos (sobre o Galnet mellorado cos candidatos validados e ampliado con variantes procedentes doutros experimentos) aumente a precisión dos seus resultados. Neste sentido, cómpre ter en conta tamén que o uso de filtros constituídos coas candidaturas que se rexeitan na revisión repercute nun melloramento inmediato da precisión da extracción nos ciclos posteriores de aplicación do experimento.

Así mesmo, queremos apuntar as posibilidades de aplicar a mesma metodoloxía para a expansión de wordnets doutras linguas que dispoñan dun recurso asimilábel a un dicionario de sinónimos.

Como liña futura de investigación, pretendemos experimentar coas posibilidades de ampliación do dicionario de sinónimos mediante técnicas de extracción léxica a partir de WordNet; é dicir, seguir a vía inversa do traballo que se presenta neste artigo.

Referencias

Gómez Clemente, Xosé María, Xavier Gómez Guinovart, Andrea González Pereira, e Verónica Taboada Lorenzo. 2013. Sinonimia e rexistros na construción do WordNet do galego. *Estudos de Lingüística Galega*, 5:27–42.

Gómez Clemente, Xosé María, Xavier Gómez Guinovart, e Alberto Simões. 2014. *Dicionario de sinónimos do galego*. Área de Normalización Lingüística, Universidade de Vigo, Vigo. URL: `http://sli.uvigo.es/sinonimos/`.

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

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

Gómez Guinovart, Xavier e Alberto Simões. 2013. Retreading dictionaries for the 21st century. En José Paulo Leal, Ricardo Rocha, e Alberto Simões, editores, *2nd Symposium on Languages, Applications and Technologies*, pp. 115–126, Saarbrücken. Dagstuhl Publishing.

Gonçalo Oliveira, Hugo. 2013. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. Tese de doutoramento, Universidade de Coimbra. URL: `http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_PhdThesis2012.pdf`.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2014. ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393.

González-Agirre, Aitor, Egoitz Laparra, e German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. En *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*, Matsue, Japan.

González-Agirre, Aitor e German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.

Miller, George A., 1998. *Nouns in WordNet*, pp. 23–46. The MIT Press, Cambridge, Massachusetts.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, e Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Solla Portela, Miguel Anxo e Xavier Gómez Guinovart. 2014. Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos. En L. Alfonso Ureña López et al., editor, *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal*, volume 1199, pp. 29–32, Aachen. CEUR Workshop Proceedings (CEUR-WS.org).