Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 119-128 recibido 04-05-12 revisado 27-05-12 aceptado 05-06-12

Revisión de técnicas para la construcción de WordNets mediante estrategia de expansión*

A revision of techniques for WordNet construction following the expand model

Antoni Oliver, Salvador Climent, Marta Contreras
Universitat Oberta de Catalunya
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,scliment,mcontrerasf@uoc.edu

Resumen: Este artículo ofrece una revisión de métodos para la construcción de WordNets siguiendo la estrategia de expansión, es decir, mediante la traducción de las *variants* inglesas del Princeton WordNet. En el proceso de construcción se han utilizado recursos libres disponibles en Internet. El artículo presenta también los resultados de la evaluación de las técnicas en la construcción de los WordNets 3.0 para el castellano y catalán. Estas técnicas se pueden utilizar para la construcción de WordNets para otras lenguas.
Palabras clave: WordNet, recursos léxicos, semántica

Abstract: This paper presents a review of methods for building WordNets following the expand model, that is, by translating the English variants of the Princeton WordNet. Only free resources available online have been used. The paper also presents the evaluation of the techniques applied in the construction of Spanish and Catalan WordNets 3.0. These techniques can be also used for other languages.
Keywords: WordNet, lexical resources, semantics

1. Introducción

WordNet (Fellbaum, 1998) es una base de datos léxica del inglés donde las palabras pertenecientes a categorías abiertas (substantivos, verbos, adjetivos y adverbios) se organizan en conjuntos de sinónimos denominados *synsets*. El WordNet original se construyó para el inglés en la Universidad de Princeton (en el resto del artículo lo denominaremos PWN - *Princeton WordNet*). Siguiendo el modelo de PWN se han construido WordNets para muchas otras lenguas. Vossen (1998) distingue dos estrategias generales para la construcción de WordNets: (1) estrategia de combinación (*merge model*), en la que se genera una ontología propia para la lengua de llegada y posteriormente se generan las relaciones con PWN; y (2) estrategia de expansión (*expand model*) en la que se traducen las *variants* en inglés utilizando diversas estrategias. En esta segunda estrategia no es necesario establecer relaciones interlingüísticas.

La principal dificultad para la construcción de WordNets mediante la estrategia de expansión es la polisemia. Si todas las *variants* fuesen monosémicas (es decir, que estuviesen asignadas a un único *synset*) el problema sería simple, ya que únicamente tendríamos que encontrar una o más traducciones para la *variant* inglesa. Al tener la palabra inglesa un único sentido la traducción de ésta seria la *variant* correcta en la lengua de llegada.

En la tabla 1 podemos observar el número de *variants* que tienen asignadas un número determinado de *synsets*. Las *variants* que tienen asignado un único *synset*, son palabras monosémicas en inglés (al menos, según PWN). Así, el 82.32 % de las *variants* del PWN son monosémicas.

Otro aspecto interesante es obser-

* Este trabajo se ha llevado a cabo dentro del proyecto Know2 *Language understanding technologies for multilingual domain-oriented information access* (MICINN, TINN2009-14715-C04-04)

Núm. synsets	variants	%
1	123.228	82.32
2	15.577	10.41
3	5.027	3.36
4	2.199	1.47
5+	3.659	2.44

Tabla 1: Número de *variants* que tienen asignados un número determinado de *synsets*

var cuántas de estas *variants* monosémicas están escritas con alguna letra en mayúsculas (y corresponderán probablemente a nombres propios) y cuántas están en minúscula. En la tabla 2 podemos observar estos valores

	variants	%
minúscula	84.714	68.75
mayúscula	38.514	31.25

Tabla 2: Número de *variants* monosémicas del PWN según se encuentren en minúculas o con alguna letra en mayúsculas

Nuestro proyecto se enmarca en la creación de los WordNets 3.0 del español y catalán. Las versiones anteriores de los WordNets en castellano y catalán diferían en su licencia: mientras que el catalán disponía de una licencia libre, el castellano se distribuía con una licencia propietaria. Los WordNets 3.0, tanto para el castellano como para el catalán, se distribuirán bajo una licencia libre.

2. *La construcción de los WordNets del español y el catalán*

Tanto el WordNet castellano (Atserias et al., 1997) como el catalán (Benítez et al., 1998) se construyeron utilizando la estrategia de expansión y de una manera prácticamente idéntica. En primer lugar se desarrollaron manualmente una serie de conceptos considerados base. Estos conceptos son los considerados más importantes y que además son comunes en todas las lenguas involucradas en el proyecto EuroWordNet (Vossen, 1999).

Para desarrollar la parte correspondiente a los substativos de las primeras versiones de estos WordNets se utilizó una metodología basada en la consulta a diccionarios bilingües. Para llevar a cabo la evaluación de este método se consideraron dos tipos de relaciones: la relación palabra inglesa - *synset*, dada por el PWN; y la relación palabra castellana[1] - palabra inglesa, dada por los diccionarios bilingües utilizados.

Atendiendo a la relación palabra inglesa - *synset* se consideraron dos grupos:

- Las palabras inglesas monosémicas, es decir, las asignadas a un único *synset*.
- Las palabras inglesas polisémicas, es decir, las asignadas a más de un *synset*

Por otro lado, atendiendo a la relación palabra castellana - palabra inglesa, se distinguieron cuatro grupos:

- Grupo 1: palabras castellanas que tienen una única traducción a una palabra inglesa, y que a su vez esta palabra inglesa sólo tiene una traducción a la misma palabra castellana.
- Grupo 2: palabras castellanas que tienen más de una traducción a diversas palabras inglesas, pero todas estas palabras inglesas se traducen únicamente por dicha palabra castellana.
- Grupo 3: palabras castellanas que se traducen por una única palabra inglesa, pero esta palabra inglesa se traduce a más de una palabra castellana.
- Grupo 4: palabras castellanas que se traducen por más de una palabra inglesa, y a su vez, cada una de estas palabras inglesas se traducen a más de una palabra castellana.

Como resultado, se obtuvieron 8 grupos de conjuntos de tres elementos (palabra castellana - palabra inglesa -*synset*). La asignación de palabra castellana a *synset* se realizó directamente en todos los grupos, pero la división permitió realizar

[1] o catalana

una evaluación para cada grupo y asignar las precisiones obtenidas como *scores* a las *variants* castellanas.

La parte verbal se desarrolló siguiendo una metodología totalmente manual dada la gran polisemia verbal y el número relativamente pequeño de conceptos verbales. Los adjetivos y adverbios no se desarrollaron en las primeras versiones de los WordNets del español y del catalán.

3. Babelnet

El objetivo de Babelnet (Navigli y Ponzetto, 2010) es crear un red semántica de grandes dimensiones combinando el conocimiento lexicográfico de WordNet con el conocimiento enciclopédico de la Wikipedia. Así pues BabelNet ofrece una relación entre las entradas de la Wikipedia y los *synsets* de WordNet. A continuación podemos observar un fragmento del archivo que se distribuye con el proyecto BabelNet:

```
Adobe_brick adobe_brick%1:06:00::
02681392n
Fuselage fuselage%1:06:00:: 03408054n
Hearse hearse%1:06:00:: 03506880n
Merida_(Yucatan) merida%1:15:00::
08740367n
```

Para poder relacionar las dos fuentes toman de WordNet todos los posibles sentidos de una determinada palabra y todas las relaciones semánticas de los *synsets*. De la Wikipedia toman todas las entradas y las relaciones dadas por los enlaces de hipertexto de las páginas. Estas relaciones pueden ser de diferentes tipos y no están especificadas desde el punto de vista semántico. Para establecer un *mapping* entre los dos recursos utilizan lo que los autores denominan *contextos de desambiguación*. Estos contextos, para los artículos de la Wikipedia están formados por las etiquetas de sentido que tienen algunas de las entradas, los enlaces de hipertexto y las categorías. En el caso de WordNet estos contextos están formados por todos los sinónimos, hiperónimos e hipónimos, los lemas de las categorías abiertas de la glosa y las *variants* asociadas a los *synsets* hermanos, es decir, los que tienen un hiperónimo directo en común. Para establecer los *mappings* aplican los siguientes criterios:

- Para todas las páginas de la Wikipedia que tengan un título monosémico tanto en la Wikipedia como en WordNet, se enlaza directamente la página con el *synset*.
- Para el resto de página se calcula la intersección de los contextos de desambiguación para todos los sentidos de la Wikipedia y de WordNet.

Dado que las páginas de la Wikipedia disponen de enlaces interlingüísticos, esta relación se puede establecer para todas las lenguas que dispongan de la entrada correspondiente.

4. Uso de corpus paralelos

Se han publicado diversos trabajos sobre el uso de corpus paralelos en tareas relacionadas con la creación de WordNets y de otras ontologías similares. En (Kazakov y Shahid, 2009) se describe una metodología para la adquisición de *variants* asociadas a *synsets* a partir de un corpus paralelo multilingüe. Las *variants* se obtienen comparando las palabras alineadas en diversas lenguas. Si una determinada palabra en una determinada lengua se traduce por más de una palabra en varias lenguas distintas, probablemente quiere decir que la palabra dada tiene más de un significado. Esta suposición funciona también al revés. Si dos palabras de una determinada lengua se traducen por una única palabra en algunas otras lenguas, probablemente quiere decir que las dos palabras comparten un significado común. En Ide, Erjavec, y Tufis (2002) se presenta una idea similar junto a una implementación práctica.

En Fišer (2007) se presenta la construcción del WordNet esloveno utilizando un corpus paralelo multilingüe, un algoritmo de alineación de palabras y WordNets existentes para otras lenguas. Se obtiene un diccionario multilingüe mediante el algoritmo de alineación de palabras y se asignan todos los *synsets* de los WordNets disponibles. Algunas de las palabras en alguna de las lenguas son polisémicas de manera que tienen asignadas más de un

synset. En algunos casos, una palabra es monosémica en al menos una de las lenguas y por lo tanto tiene un único *synset* asignado. Este *synset* se utiliza para desambiguar y asignar un único *synset* al resto de lenguas, incluído el Esloveno. En Sagot y Fišer (2008) se utiliza un método muy parecido para el francés, junto a métodos basados en diccionarios.

En algunos trabajos anteriores presentamos una metodología de construcción de WordNets basadas en el uso de corpus bilingües paralelos. Estos corpus necesitaban estar anotados semánticamente en la parte inglesa. Como este tipo de corpus no está fácilmente disponible exploramos dos estrategias distintas para poder construir automáticamente los corpus necesarios:

- Mediante traducción automática de corpus anotados semánticamente (Oliver y Climent, 2011), (Oliver y Climent, 2012a)
- Mediante anotación semántica automática de corpus bilingües paralelos (Oliver y Climent, 2012b)

5. Estrategias evaluadas

En nuestos experimentos, hemos utilizado y evaluado tres estrategias distintas:

- Uso de diccionarios bilingües
  - Diccionarios bilingües generales
  - Diccionarios enciclopédicos
  - Diccionarios terminológicos
- Uso de Babelnet
- Uso de corpus paralelos

La primera estrategia, la basada en diccionarios bilingües la dividimos en tres grupos, dependiendo del tipo de diccionario utilizado. En esta primera estrategia obtenemos *variants* únicamente para *synsets* cuyas *variants* en inglés son monosémicas. Es decir, traducimos mediante diferentes tipos de diccionarios (generales, enciclopédicos o terminológicos) palabras inglesas monosémicas (asignadas a un único *synset*) y asignamos este *synset* a la correspondiente palabra o palabras castellanas o catalanas dadas por el diccionario. Es decir, tratamos los cuatro grupos monosémicos comentados en la sección 2 simultáneamente. Las dos últimas estrategias tratan tanto los casos monosémicos como polisémicos.

En todos los casos utilizamos una evaluación automática, comparando los resultados obtenidos con las versiones preliminares de los WordNets 3.0 para el castellano y catalán. De esta manera, no todos los resultados obtenidos pueden ser evaluados, ya que algunos *synsets* no tienen *variants* asignadas en estas versiones preliminares. Las *variants* obtenidas para un determinado *synset* que coincidan con las *variants* registradas en la versión preliminar para el mismo *synset* serán consideradas correctas. Si la *variant* obtenida no coincide con ninguna de las registradas en la versión preliminar, será considerada incorrecta. Si no tenemos registrada ninguna *variant* para el *synset* en cuestión, no se contabilizará para el cálculo de la precisión. Hay que tener en cuenta que los resultados considerados errores por la evaluación automática pueden ser correctos en realidad y que simplemente se trate de una *variant* no registrada en la versión preliminar. La evaluación de cada experimento nos produce tres listas: los resultados correctos, que en principio no requieren verificación posterior; los resultados incorrectos, que pueden no serlo y que en algunos casos se han verificado manualmente; los resultados que no se han podido evaluar y que no intervienen en el cálculo de la precisión calculada automáticamente. En algunos experimentos también se ha revisado manualmente estos últimos resultados.

En los próximos apartados presentamos las tres estrategias utilizadas y sus evaluaciones.

6. Uso de diccionarios bilingües

6.1. Uso de diccionarios bilingües generales

6.1.1. Metodología

En este experimento hemos utilizado un diccionario bilingüe para traducir las *variants* inglesas monosémicas escritas minúsculas en el PWN 3.0. Como podemos observar en la tabla 2 éstas son el

68.75 % de las *variants* inglesas monosémicas. De éstas, la mayoría corresponden a substantivos (74.26 %), seguidas de adjetivos (16.48 %). En mucho menor porcentaje se encuentran verbos (5.35 %) y adverbios (3.91 %).

6.1.2. Recusos utilizados

El único recurso necesario es un diccionario bilingüe que tenga una buena cobertura. En nuestros experimentos hemos obtenidos diccionarios bilingües inglés-castellano e inglés-catalán a partir de los diccionarios de transferencia de Apertium (Forcada, Tyers, y Ramírez-Sánchez, 2009) y a partir del Wiktionary[2]. En la tabla 3 podemos observar el número de entradas de cada uno de los diccionarios, así como el número de entradas una vez combinados ambos diccionarios para cada par de lenguas.

Diccionario	**eng-spa**	**eng-cat**
Apertium	20.366	29.154
Wiktionary	23.196	7.393
Total	34.600	32.921

Tabla 3: Número de entradas de los diccionarios bilingües.

6.1.3. Evaluación

Para el castellano podemos obtener un total de 12.676 *variants* de las cuales 7.401 son correctas, 2.997 incorrectas (según la evaluación automática) y no podemos evaluar 2.278. La precisión para el castellano, según la evaluación automática, es del 71.2 %. Se han revisado manualmente todos los resultados considerados incorrectos por la evaluación automática. Esto nos ha permitido calcular una nueva precisión, que ahora asciende al 93.95 %.

Para el catalán obtenemos un total de 8.335 *variants*, de las cuales 4.223 son correctas, 1.083 incorrectas (según la evaluación automática) y no podemos evaluar 3.029. La precisión para el catalán, según la evaluación automática, es del 79.6 %. Del mismo modo que para el castellano, hemos revisado los resultados incorrectos y hemos podido calcular una nueva precisión que asciende al 96.36 %.

[2]http://www.wiktionary.org

6.2. Uso de diccionarios enciclopédicos

6.2.1. Metodología

En este experimento se ha utilizado un diccionario enciclopédico para traducir las *variants* inglesas monosémicas escritas con la primera letra en mayúscula. Éstas constituyen el 31.25 % de las *variants* monosémicas, como se puede ver en la tabla 2. De estos la inmensa mayoría (99.17 %) son substantivos.

6.2.2. Recusos utilizados

Se ha creado un diccionario bilingüe inglés-castellano y uno inglés-catalán a partir de la Wikipedia inglesa, utilizando los enlaces interlingüísticos. De esta manera se han obtenido 59.659 entradas para el inglés-castellano y 22.205 entradas para el inglés-catalán.

6.2.3. Evaluación

Para el castellano podemos obtener un total de 10.356 *variants* de las cuales 4.722 son correctas, 1.916 incorrectas (según la evaluación automática) y no podemos evaluar 3.718. La precisión para el castellano, según la evaluación automática, es del 71.1 %. Si revisamos los casos dados por incorrectos y recalculamos la precisión, ésta sube hasta el 89.74 %. Algunos casos incorrectos lo son por pequeños detalles tipográficos fácilmente subsanables manualmente. Si consideramos este último grupo también como correcto, la precisión asciende hasta el 90.37 %.

Para el catalán obtenemos un total de 7.083 *variants*, de las cuales 2.642 son correctas, 1.278 incorrectas (según la evaluación automática) y no podemos evaluar 3.163. La precisión para el catalán, según la evaluación automática, es del 67.4 %. Después de la revisión manual de los clasificados como incorrectos, la precisión aumenta hasta el 90.94 %, y considerando los que requieren poca modificación manual ésta asciende hasta el 98.34 %.

6.3. Uso de diccionarios terminológicos

6.3.1. Metodología

En este experimento hemos utilizado un conjunto de diccionarios terminológicos multilingües para traducir las *variants*

inglesas monosémicas, tanto las que están en minúsculas como las que están en mayúsculas, ya que los diccionarios terminológicos incluyen a ambas.

6.3.2. Recusos utilizados

El único recurso necesario es un diccionario terminológico que contenga las lenguas que nos interesan y que tenga una buena cobertura. En nuestros experimentos hemos obtenido diccionarios terminológicos bilingües inglés-castellano e inglés-catalán a partir de todos los diccionarios terminológicos que ofrece Termcat en su apartado Terminologia Oberta[3]. De esta manera se ha confeccionado un diccionario terminológico inglés-castellano de 46.761 entradas y uno inglés-catalán de 46.653 entradas.

6.3.3. Evaluación

Para el castellano obtenemos un total de 10.456 *variants*, de las cuales 4.180 son correctas, 3.346 incorrectas (según la evaluación automática) y no podemos evaluar 2.930. La precisión para el castellano, según la evaluación automática, es del 55.5 %. Este resultado es muy bajo, por lo que decidimos revisar manualmente tanto las evaluadas automáticamente como incorrectas, como las no evaluadas. Muchas de las evaluadas automáticamente como incorrectas eran en realidad correctas y muchas de las no evaluadas también lo eran. Con estos nuevos datos, hemos calculado la precisión que es ahora del 98.57 %.

Para el catalán podemos obtener un total de 9.890 *variants* de las cuales 3.007 son correctas, 2.614 incorrectas (según la evaluación automática) y no podemos evaluar 4.269. La precisión para el catalán calculada automáticamente es del 53.5 %. Procedemos de igual manera que para el castellano y obtenemos una nueva precisión del 98.36 %.

7. Uso de Babelnet

7.1. Metodología

Para obtener los WordNets español y catalán a partir de BabelNet utilizaremos el fichero babel-to-wordnet-3.0.txt cuyo contenido hemos mostrado en la sección 3. Este fichero relaciona los *synsets* de WordNet con sus correspondientes entradas de la Wikipedia. Los títulos de las entradas serán las *variants* inglesas. A partir de los enlaces interlingüísticos de la Wikipedia podemos relacionar los títulos de las entradas inglesas con los de la española y catalana, obteniendo de este modo las correspondientes *variants* españolas y catalanas. De hecho, en este experimento utilizamos también los diccionarios enciclopédicos descritos en 6.2, ya que éstos se han obtenido a partir de los enlaces interlingüísticos de la Wikipedia.

7.2. Recusos utilizados

Como ya hemos comentado, utilizamos el fichero babel-to-wordnet-3.0.txt del proyecto BabelNet y los diccionarios enciclopédicos descritos en 6.2.

7.3. Evaluación

Para el castellano obtenemos un total de 26.209 *variants*, de las cuales 14.614 son correctas, 5.065 incorrectas (según la evaluación automática) y no podemos evaluar 6.530. La precisión para el castellano, según la evaluación automática, es del 74.3 %. Revisamos manualmente tanto las evaluadas automáticamente como incorrectas, como las no evaluadas. Una vez realizada la revisión manual podemos calcular un nuevo valor de precisión que es de 81.02 %. Si observamos la revisión manual, nos damos cuenta que una cantidad considerable de errores lo son por pequeños detalles ortotipográficos o por faltar o sobrar algún carácter. Si consideramos estos casos como correctos, la precisión aumenta hasta el 89.24 %.

Para el catalán podemos obtener un total de 18.366 *variants* de las cuales 9.044 son correctas, 3.548 incorrectas (según la evaluación automática) y no podemos evaluar 5.774. La precisión para el catalán, según la evaluación automática, es del 61 %. Hemos procedido de la misma manera que para el castellano y hemos calculado una nueva precisión del 80.91 %, que sube hasta el 97.43 % si consideramos correctas las que han sufrido pequeñas modificaciones.

[3] http://www.termcat.cat/productes/toberta.htm

8. Uso de corpus paralelos

8.1. Metodología

La metodología que presentamos se basa en la alineación a nivel de palabras de un corpus paralelo formado por el original en inglés etiquetado semánticamente y la traducción al castellano o catalán. En este artículo presentaremos únicamente los resultados obtenidos para el castellano. El corpus en inglés está etiquetado semánticamente utilizando los *synsets* de WordNet como etiquetas. En realidad la alineación que nos interesa es la alineación *synset* - palabra castellana, que es directamente deducible del corpus. A continuación podemos observar un ejemplo:

```
English:
Then he noticed that the dry wood of the wheels
had swollen.
Sense Tagged English:
00117620r he 02154508v that the 02551380a
15098161n of the 04574999n had 00256507v .
Spanish Translation:
Entonces se dio cuenta de que la madera seca
de las ruedas se había hinchado.
```

De este fragmento del corpus podremos obtener las siguientes relaciones mediante el algoritmo de alineación de palabras:

```
00117620r - entonces
02154508v - darse cuenta
02551380a - seco
15098161n - madera
```

8.2. Recusos utilizados

Como hemos comentado, exploramos dos estrategias para obtener el corpus necesario. Una se basa en la traducción automática de corpus etiquetados semánticamente. En este caso utilizamos el corpus Semcor[4] (Miller et al., 1993) y el Princeton WorNet Gloss Corpus (PWGC)[5] y el traductor automático Google Translate[6]. La segunda estrategia se basa en el etiquetado semántico automático de corpus paralelos. En este caso utilizamos en corpus European Parliament Proceedings Parallel Corpus (Europarl)[7] (Koehn, 2005) y Freeling (Padró et al., 2010) como etiquetador. En todos los casos se ha utilizado el Berkeley Aligner (Liang, Taskar, y Klein, 2006) como algoritmo para la alineación de palabras.

[4]http://www.cse.unt.edu/∽rada/downloads.html
[5]http://wordnet.princeton.edu/glosstag.shtml
[6]http://translate.google.com
[7]http://www.statmt.org/europarl/

8.3. Evaluación

La evaluación en este caso se ha llevado a cabo de manera automática y de una manera acumulativa empezando por el *synset* más frecuente en el corpus y siguiendo un orden descendiente en frecuencia. Los resultados se presentan de manera gráfica donde los valores del eje y representan la precisión acumulada y los valores del eje x el número de *synsets* extraídos (es decir, el número de *variants* asociadas a *synsets*).

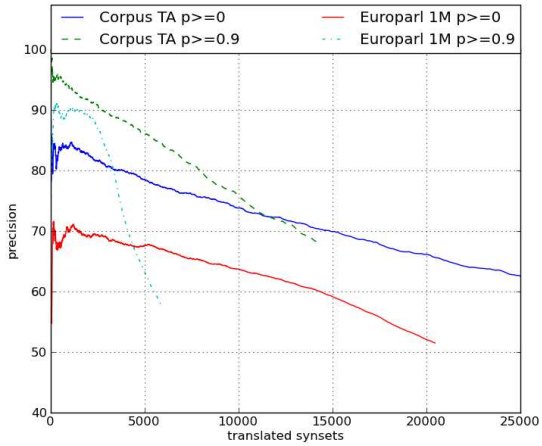

Figura 1: Comparación de los métodos basados en corpus paralelos.

En la figura 1 podemos observar los valores correspondientes al uso del corpus obtenido por traducción automática (Corpus TA) y los correspondientes a un fragmento de 1 millón de oraciones del corpus Europarl. Para ambos corpus ofrecemos dos gráficos, uno tomando todas las posibles alineaciones ($p \geq 0$) y otro tomando únicamente las alineaciones que tienen una probabilidad de 0.9 o superior ($p \geq 0{,}9$). Los mejores resultados se obtienen para el corpus obtenido mediante traducción automática y para $p \geq 0{,}9$, ya que podemos obtener unas 5.000 *variants* con una precisión del 85 % o superior.

9. Conclusiones

En este artículo hemos presentado una revisión de métodos de construcción de WordNets mediante la estrategia de expansión. Hemos presentado valores de evaluación de estos métodos para el castellano y catalán. Los valores obtenidos son similares a los obtenidos por Atserias et al. (1997) y Benítez et al. (1998). Los valores de precisión de nuestros métodos basados en diccionarios son similares a los valores obtenidos por los cuatro criterios monosémicos (que son de entre el 85 % y el 92 % para el castellano y entre el 93.3 % al 97.6 % para el catalán). La estrategia basada en BabelNet y en corpus paralelos no son fácilmente comparables con los resultados de Atserias y Benítez ya que no hemos realizado una separación de los resultados según el grado de monosemia-polisemia.

Un aspecto importante que se deduce del artículo es que la evaluación automática no es suficientemente fiable. En algunos casos hemos pasado de precisiones del 53.5 % al 98.36 % una vez revisados manualmente todas las *variants* obtenidas.

Así pues, los métodos presentados se pueden considerar de gran utilidad para la creación de WordNets. La validación final de los resultados es importante, pero esta validación es mucho más rápida que la creación de WordNets manualmente desde cero.

Como trabajo futuro pretendemos mejorar los resultados obtenidos mediante el uso de corpus paralelos. Por un lado, si utilizamos corpus etiquetados semánticamente y traducción automática, tenemos un problema de precisión dado por la calidad de la traducción automática (especialmente en la tarea de selección léxica) y un problema de cobertura ya que en el extremo podremos obtener únicamente las *variants* correspondientes a los *synsets* presentes en el corpus. Por este motivo, y dada la relativa facilidad de obtener corpus paralelos multilingües de gran tamaño, preferimos trabajar con esta segunda estrategia. En este caso, la precisión de la tarea de etiquetado semántico automático no es suficiente. Por este motivo queremos aprovechar la información presente en los corpus paralelos multilingües para llevar a cabo una desambiguación a través del etiquetado de diversas lenguas que dispongan de WordNets, tal y como se explica en (Ide, Erjavec, y Tufis, 2002).

Por otro lado, en los experimentos llevados a cabo hasta el momento con corpus paralelos hemos extraído los WordNets partiendo de cero. Como que ya disponemos de versiones de los WordNets para el castellano y catalán, podemos utilizar toda esta información para simplificar la tarea de extracción de nuevas *variants*.

Los WordNets 3.0 para el castellano y catalán desarrollados en este proyecto se pueden descargar libremente de http://adimen.si.ehu.es/web/MCR . Los WordNets del castellano y catalán tienen una licencia Creative Commons, concretamente la Attribution 3.0 Unported (CC BY 3.0) license[8].

Bibliografía

Atserias, J., S. Climent, X. Farreres, G. Rigau, y H. Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingual WordNets. En *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volumen 97, página 327–338.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, y Mariona Taulé. 1998. Methods and tools for building the catalan WordNet. En *In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages*.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. The MIT press.

Fišer, D. 2007. Leveraging parallel corpora and existing wordnets for automatic construction of the slovene wordnet. En *Proceedings of the 3rd Language and Technology Conference*, volumen 7, página 3–5.

Forcada, M. L, F. M Tyers, y G.írez-Sánchez@. 2009. The apertium machine translation platform: five years on.

[8] http://creativecommons.org/licenses/by/3.0/

En *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, página 3–10.

Ide, N., T. Erjavec, y D. Tufis. 2002. Sense discrimination with parallel corpora. En *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, página 61–66.

Kazakov, D. y A.R. Shahid. 2009. Unsupervised construction of a multilingual WordNet from parallel corpora. En *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, página 9–12.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. En *MT summit*, volumen 5.

Liang, Percy, Ben Taskar, y Dan Klein. 2006. Alignment by agreement. En *Proceedings of the HLT-NAACL '06.*

Miller, George A, Claudia Leacock, Randee Tengi, y Ross T Bunker. 1993. A semantic concordance. En *Proceedings of the workshop on Human Language Technology*, HLT '93, página 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1075742.

Navigli, Roberto y Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, página 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.

Oliver, A. y S. Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. En *Proceedings of the 27th Conference of the SEPLN, Huelva Spain.*

Oliver, A. y S. Climent. 2012a. Building wordnets by machine translation of sense tagged corpora. En *Proceedings of the Global WordNet Conference, Matsue, Japan.*

Oliver, A. y S. Climent. 2012b. Parallel corpora for wordnet construction. En *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2012). New Delhi (India).*

Padró, L., S. Reese, E. Agirre, y A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and UKB. En *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010).*

Sagot, B. y D. Fišer. 2008. Building a free french wordnet from multilingual resources. En *Proceedings of OntoLex 2008*, Marrackech (Morocco).

Vossen, P. 1998. Introduction to Eurowordnet. *Computers and the Humanities*, 32(2):73–89.

Vossen, P. 1999. EuroWordNet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).