Construcción de los WordNets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente

Antoni Oliver, Salvador Climent
Universitat Oberta de Catalunya (UOC)
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,scliment@uoc.edu

28 de junio de 2011

Resumen

Este artículo describe una metodología de construcción de WordNets que se basa en la traducción automática de un corpus en inglés desambiguado por sentidos. El corpus que utilizamos está formado por las propias glosas de WN 3.0 etiquetadas semánticamente y por el corpus Semcor. Los resultados de precisión son comparables a los obtenidos mediante métodos basados en diccionarios bilingües para las mismas lenguas. La metodología descrita se está utilizando, en combinación con otras estrategias, en la creación de los WordNets 3.0 del español y catalán.

Building Catalan and Spanish WordNet 3.0 by machine translation of sense disambiguated corpora

This paper describes a methodology for the construction of WordNets based on machine translation of an English sense tagged corpus. We use the Semcor corpus and the WordNet 3.0 sense tagged glosses as a corpus. Precision results are comparable to those obtained by methods based on bilingual dictionaries for the same languages. This methodology is being used for the construction of the Spanish and Catalan WordNets 3.0 in combination with other strategies

1. Introducción

WordNet (Fellbaum, 1998) es una base de conocimiento léxico del inglés que organiza los sustantivos, verbos, adjetivos y adverbios en conjuntos de sinónimos que reciben el nombre de *synsets*. Cada *synset* representa un concepto lexicalizado en inglés, y se conecta a los otros *synsets* mediante relaciones semánticas (hiponimia, antonimia, meronimia, troponimia etc.). Por ejemplo, el synset del WordNet 3.0 que se identifica con el *offset* i la categoria gramatical 06171040-n, tiene asignada dos *variants* (o sinónimos): *linguistics* y *philology*. Cada *synset* tiene asignada también una glosa o definición, que en el ejemplo que nos ocupa es: *the humanistic study of language and literature*. También podemos obtener relaciones con otros *synsets*, por ejemplo, tiene un hiperónimo: 06153846-n (*humanistic_dicipline*, *humanities*); y dos hipónimos 06171265-n (*dialectology*) i 06178812-n (*lexicology*). WordNet se ha convertido en un recurso estándar en todo tipo de investigaciones y aplicaciones semánticas en el área del Procesamiento del Lenguaje Natural.

El WordNet inglés es libre y se puede descargar de su página web de la Universidad de Princeton[1] (en el resto del artículo denominaremos a este WordNet PWN - *Princeton WordNet*). La versión actual es la 3.0, lanzada en diciembre de 2006. En la tabla 1 podemos observar una comparativa del número de *synsets* para las versiones 1.5, 1.6 y 3.0 del PWN. Como podemos ver, el número de *synsets* desarrollados aumenta con las nuevas versiones, lo que obliga a desarrollar nuevos *synsets* en los wordnets para otras lenguas con el objetivo de lograr que los recursos sean comparables.

	1.5	**1.6**	**3.0**
Total	76.705	99.642	118.695
Sustantivos	51.253	66.025	83.073
Verbos	8.847	12.127	13.845
Adjetivos	13.460	17.915	18.156
Adverbios	3.145	3.375	3.621

Cuadro 1: Comparación del número de synsets para tres versiones del PWN

[1] http://wordnet.princeton.edu/

Diversos proyectos han desarrollado wordnets para otras lenguas: EuroWordNet (Vossen, 1998) inicialmente para holandés, italiano, español y la expansión y mejora del inglés, y en una extensión del proyecto para alemán, francés, estonio y checo; Balkanet (Tufis, Cristea, y Stamou, 2004) para búlgaro, griego, rumano, serbio y turco y RusNet (Azarova et al., 2002) para el ruso, entre otros. En la página web de la Global WordNet Association[2] podemos ver una lista exhaustiva de los WordNets disponibles para diversas lenguas.

En la tabla 2 podemos ver los WordNets que, según nuestro conocimiento, se distribuyen bajo una licencia libre y el número de synsets que contienen[3].

Muchos de los WordNets disponibles están sujetos a licencias comerciales. Nuestro trabajo se engloba en un proyecto cuyo objetivo es desarrollar la versión 3.0 para el castellano y catalán y distribuirla con una licencia libre (GNU GPL).

Para ello se aplicará una metodología basada en la traducción automática, que se describe con detalle en 3. En 4 presentamos la evaluación de los métodos y finalmente en 5 las conclusiones y el trabajo futuro.

2. Estrategias de construcción de WordNets

En esta sección repasaremos algunas técnicas que se han utilizado para construir WordNets para diversas lenguas distintas del inglés. Se pueden distinguir dos aproximaciones generales para la construcción de WordNets (Vossen, 1998):

- **Estrategia de combinación** (*merge model*): se genera una ontología propia para la lengua de llegada, con sus propios niveles y relaciones. Posteriormente se generan las relaciones interlingüísticas de este WordNet con PWN.

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

[3] El número de synsets puede ser diferente del especificado, ya que algunos de estos proyectos todavía están activos. Para el ruso no hemos podido determinar el número de synsets.

Lengua	**Synsets**
Inglés	117.775
Catalán	100.442
Thai	73.595
Danés	59.000
Japonés	57.241
Coreano	42274
Tamil	41.442
Lao	38.886
Hindi	34.114
Indonesio	33.721
Irlandés	32.742
Francés	31.822
Esloveno	20.000
Burmés	19.943
Vietnamita	12.271
Hebreo	5.922
Albanés	4.688
Mongol	1.623
Bengalí	635
Sinhala	270
Sundanés	70
Nepalí	47
Ruso	?

Cuadro 2: Lenguas que disponen de WordNets con licencia libre

- **Estrategia de expansión** (*expand model*): se traducen las *variants* asociadas a los *synsets* de PWN, utilizando diccionarios bilingües u otras estrategias. Posteriormente se revisa si las relaciones entre *synsets* impuestas por la estructura de PWN son válidas para la lengua de llegada. No es necesario establecer relaciones interlingüísticas porque el WordNet local y PWN son paralelos.

Cada una de estas estrategias presenta una serie de ventajas e inconvenientes (Vossen, 1996). Por un lado, la estrategia de expansión es más sencilla desde un punto de vista técnico y garantiza un grado mayor de compatibilidad entre los WordNets de las diferentes lenguas. Por contra, los WordNets

desarrollados mediante esta técnica están demasiado influenciados por PWN y contendrán todos sus errores y deficiencias estructurales. La estrategia de combinación es más compleja desde el punto de vista técnico, pero permite un aprovechamiento más directo de las ontologías y tesauros existentes.

2.1. Construcción de los primeros WordNets del español y el catalán

Las respectivas construcciones de los WordNets 1.5 del español (Atserias et al., 1997) y el catalán (Benítez et al., 1998), los primeros que existieron en dichas lenguas, siguieron una metodología prácticamente idéntica, que se puede clasificar como de expansión, ya que fundamentalmente consistió en la traducción de las *variants* correspondientes a los *synsets* del PWN. En este apartado daremos algunos detalles sobre la construcción de estos WordNets, ya que estas primeras versiones son los puntos de partida de la construcción de las versiones posteriores. Nos centraremos en los conceptos nominales, ya que los verbales se desarrollaron de una manera manual y los adjetivos y adverbios no se desarrollaron en las primeras versiones.

Los conceptos nominales se han generado automáticamente a partir de diccionarios bilingües. Para evaluar los resultados se dividieron los *synsets*, palabras inglesas y palabras españolas[4] atendiendo a dos relaciones: la relación palabra inglesa - *synset* y la relación palabra española - palabra inglesa. Así, atendiendo a la relación palabra inglesa - *synset* se pueden considerar dos grupos:

- Las palabras inglesas monosémicas, que son aquellas que están asignadas a un único *synset*.
- Las palabras inglesas polisémicas, que son aquellas que están asignadas a más de un *synset*.

Atendiendo a la relación palabra española - palabra inglesa se pueden distinguir cuatro grupos:

[4]O catalanas, ya que la metodología utilizada para las dos lenguas son prácticamente idénticas

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

Como resultado se obtienen 8 grupos de tres elementos palabra española - palabra inglesa - *synset*. Adicionalmente se aplicó un criterio denominado de *variant* que enlaza directamente palabras españolas a *synsets*. Este enlace se establece en los casos en los que un *synset* tiene asignadas dos o más *variants* en PWN y estas variants tienen como traducción la misma palabra española.

En los artículos citados se pueden encontrar resultados detallados de precisión para cada uno de los grupos. En resumen, para el castellano obtenemos precisiones que van del 54.5 % para el grupo polisémico 4 al 97.6 % para el grupo monosémico 2. Para el catalán las precisiones van del 58 % para el grupo polisémico 3 al 92 % para el grupo monosémico 1.

2.2. Usos de traducción automática para la construcción de WordNets

En este apartado se muestran sucintamente dos trabajos relacionados con WordNet que utilizan traducción automática de corpus etiquetados semánticamente: la construcción del WordNet macedonio y el proyecto BabelNet que enlaza páginas de la Wikipedia con *synsets* del PWN.

En la construcción del WordNet macedonio (Saveski y Trajkovski, 2010) se ha utilizado como fuente de información principal un diccionario bilingüe inglés - macedonio. Cuando las entradas son monosémicas la relación *synset* - palabra macedonia se puede establecer directamente. En los caso de polisemia esta asignación se puede reducir a un problema de desambiguación de sentidos (WSD - *Word Sense Disambiguation*). Los autores utilizan la propuesta de (Dagan y Itai, 1994) que se basa en el hecho de que una palabra determinada tiende a coaparecer más en un corpus de gran tamaño con las palabras de categoría abierta de su propia definición que con otras palabras. En el caso del macedonio los autores se encontraban con dos problemas: (i) el diccionario utilizado no disponía de definiciones; y (ii) no existe un corpus monolingüe de gran tamaño para el macedonio. El primer problema lo solucionan traduciendo las glosas de PWN al macedonio utilizando Google Translate, de esta manera, las glosas traducidas hacen la función de definición. El segundo problema lo solucionan utilizando la web como corpus a través de la *Google Similarity Distance* (Cilibrasi y Vitanyi, 2007). Esta distancia entre una palabra o frase x y una palabra o frase y se calcula a partir de los resultados de las búsquedas a Google de x, de y y de una consulta que incluya a x e y.

El objetivo de BabelNet (Navigli y Ponzetto, 2010) es crear una red semántica de grandes dimensiones integrando el conocimiento lexicográfico de WordNet con el conocimiento enciclopédico de la Wikipedia. Una vez realizada esta integración y dado que las páginas de la Wikipedia tienen enlaces interlingüísticos, se puede establecer la relación para todas las lenguas que dispongan de la entrada dada. Así pues, si disponemos de una relación entre el *synset* s y la entrada de la wikipedia en inglés w_{eng}, mediante los enlaces interlingüísticos podemos establecer esta misma relación para el resto de lenguas que dispongan de la entrada correspondiente: w_{spa}, w_{fra}, etc. Si una lengua determinada no dispone de la entrada equivalente a w_{eng}, los autores utilizan el sistema de traducción automática Google Translate para traducir una serie de oraciones que contengan las *variants* del *synset* s. Estas oraciones las toman del corpus desambiguado Semcor (Miller et al., 1993) y de frases de la Wikipedia que contienen enlaces a la página correspondiente a w_{eng}. Una vez realizada la traducción automática, se identifica la traducción

más frecuente y se incluye en BabelNet.

3. Nuestra aproximación. Parte experimental

En nuestro trabajo presentamos y evaluamos una metodología de construcción de WordNets basada en traducción automática de corpus etiquetados semánticamente con los *synsets* de PWN 3.0. En primer lugar describiremos los recursos y herramientas utilizados y a continuación la metodología propiamente dicha.

3.1. Recursos lingüísticos

Para llevar a cabo los experimentos necesitamos disponer de un corpus lingüístico en inglés que esté etiquetado y desambiguado semánticamente. Las etiquetas utilizadas deben ser los propios *synsets* del PWN. Existen dos recursos libres que pueden cumplir perfectamente esta función:

- El corpus Semcor[5] (Miller et al., 1993).
- Las propias glosas del PWN 3.0 etiquetadas semánticamente[6].

3.2. Herramientas utilizadas

3.2.1. Sistemas de traducción automática

En este trabajo utilizamos dos sistemas de traducción automática:

- Google Translate[7] que nos permite traducir del inglés al castellano y al catalán.
- Sistema de traducción automática de Microsoft[8] que nos permite traducir del inglés al castellano.

[5]http://www.cse.unt.edu/ rada/downloads.html
[6]http://wordnet.princeton.edu/glosstag.shtml
[7]http://translate.google.com
[8]http://www.microsofttranslator.com/

Así, para el par de lenguas inglés-castellano dispondremos de dos sistemas de traducción automática, mientras que para el catalán únicamente dispondremos de uno. Ambos sistemas son estadísticos y trabajan a partir de corpus de gran tamaño.

3.2.2. Análisis morfosintáctico

Todas las oraciones del corpus, tanto originales como traducidas, se han etiquetado morfosintácticamente utilizando la librería Freeling (Padró et al., 2010a).

3.3. Metodología

La hipótesis de trabajo es que estos sistemas de traducción automática realizan una selección léxica adecuada y permiten desambiguar sentidos del original. No hemos realizado una comprobación exhaustiva de esta hipótesis, pero hemos comprobado algún caso aislado. Un posible ejemplo es la palabra inglesa *wood* que pueden significar, entre otros, madera o bosque (correspondiente a los *synsets* del PWN 15098161-n (*the hard fibrous lignified substance under the bark of trees*) y 08438533-n (*the trees and other plants in a large densely wooded area*) respectivamente). Si tomamos una frase en inglés correspondiente al primer sentido, por ejemplo:

```
This house is made of wood.
```

obtenemos las siguientes traducciones con los sistemas utilizados:

```
Google-es: Esta casa es de madera.
Micro.-es: Esta casa está hecha de madera.
Google-ca: Aquesta casa és de fusta.
```

Y si tomamos una frase correspondiente al segundo sentido:

```
He got lost in the wood beyond Seattle.
```

obtenemos las siguientes traducciones con los sistemas utilizados:

```
Google-es: Se perdió en el bosque más allá de Seattle.
Micro.-es: Se perdió en el bosque más allá de Seattle.
Google-ca: Es va perdre en el bosc més enllà de Seattle.
```

Vemos, pues, que los sistemas utilizados pueden realizar la selección léxica adecuadamente al menos en algunos casos. El método de evaluación descrito en 4 es un método indirecto de comprobación de la adecuación y efectos de la hipótesis.

Así pues, el procesamiento de los recursos con las herramientas descritas nos permite obtener:

- Corpus equivalentes en inglés, castellano y catalán etiquetados morfosintácticamente.
- Un conjunto de índices que relacionan *synsets* de PWN 3.0 con las oraciones del corpus donde aparecen.

Esto nos permite obtener para cada lengua todas las oraciones en las que aparece un determinado *synset*. Además, al estar estas oraciones etiquetadas morfosintácticamente podemos aplicar diversos algoritmos que nos permitan obtener la *variant* más probable de un determinado *synset* en cualquiera de las lenguas. A continuación presentamos dos de estos algoritmos, los cuales presentan dos limitaciones:

- Sólo detectan como *variant* unidades léxicas simples.
- Sólo detectan una *variant* para cada *synset*.

Ambas limitaciones se tratarán en posteriores versiones de los algoritmos.

3.3.1. Algoritmo A

El primer algoritmo funciona de la siguiente manera:

- Se toman todos los *synsets* presentes en el corpus ordenados por frecuencia de aparición, empezando por el más frecuente.

- Para cada *synset* se obtienen subcorpus (analizados morfosintácticamente) formados por todas las oraciones en qué éste aparece, en la lengua deseada y si es traducción, con el traductor automático que especifiquemos.

- El lema más frecuente de este subcorpus que comparta categoría gramatical con el *synset* que estamos tratando es elegido como resultado, es decir, como *variant* apropiada para el *synset*.

Por ejemplo: el *synset* animal.n.01 aparece 477 veces en el corpus inglés y una vez aplicado el algoritmo los candidatos a *variant* en castellano con sus frecuencias son:

```
animal:463;planta:76;especie:21;
forma:20;ser:19;ave:19;cuerpo:18;...
```

por lo que tomaremos *animal* como *variant* correspondiente en castellano.

Por su parte, el *synset* hold.v.01, con una frecuencia de 291 ofrece los siguientes resultados:

```
ser:27;mantener:25;tener:17;haber:16;
estar:12;celebrar:11;decir:6;unir:6;...
```

En consecuencia, en este caso tomaremos, de manera errónea *ser* como *variant* correspondiente.

3.3.2. Algoritmo B

Este algoritmo funciona exactamente igual que el A, pero se toma únicamente la propuesta más frecuente si su frecuencia es como mínimo el doble que la del siguiente candidato. En caso contrario, no se obtiene ningún resultado para el *synset*. Este margen del 100 % de distancia se ha tomado a partir de unos experimentos que han demostrado que ofrecen un buen compromiso entre precisión y cobertura. En la figura 1 se puede observar una comparativa de diversos valores de percentaje.

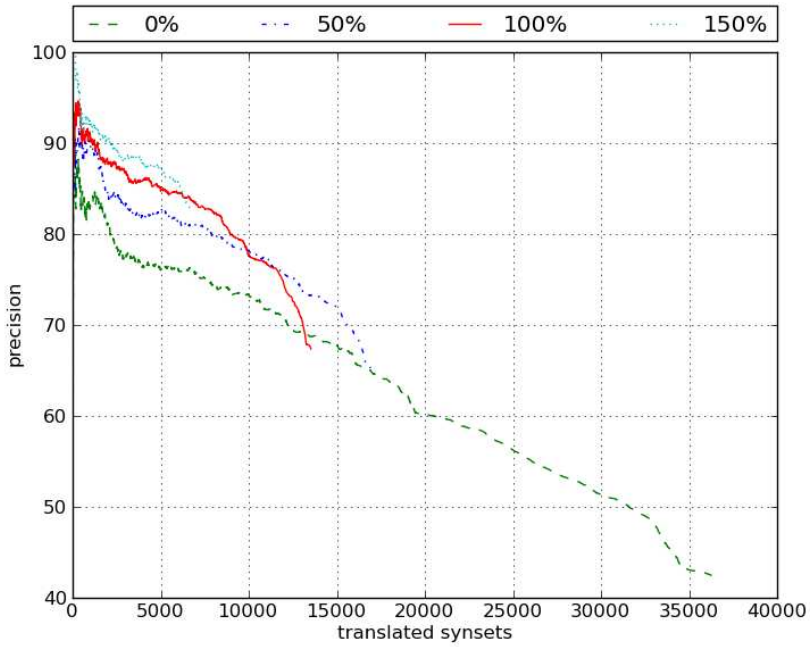

Figura 1: Comparación % de incremento de frecuencia del primer candidato respecto al segundo para el catalán.

4. Evaluación

En este apartado presentamos los resultados de la evaluación de los algoritmos presentados en la sección anterior. Cada uno de los algoritmos se evalúa independientemente y se realiza una evaluación para cada una de las tres lenguas:

- Inglés: aplicamos los algoritmos con el objetivo de valorar los resultados que se obtendrían con un sistema de traducción automática perfecto, ya que el inglés es la lengua original.
- Español: dado que disponemos de dos sistemas de traducción automática, evaluaremos los resultados para cada sistema y para una combinación de ambos. Esta combinación consiste simplemente en tomar las traducciones de ambos sistemas.
- Catalán: únicamente disponemos de un sistema de traducción automática.

La evaluación de los sistemas se ha realizado automáticamente comparando los resultados con los siguientes recursos:

- Inglés: el propio PWN 3.0.
- Español: una primera versión del WN 3.0 obtenida a partir del fragmento libre de WN 1.6 convertido a 3.0 mediante un *mapping* más un conjunto de *synsets* traducidos y validados manualmente (Fernández-Montraveta, Vázquez, y Fellbaum, 2008). En el momento de realizar este trabajo esta versión de WN 3.0 contaba con 39.434 *synsets* y 53.530 *variants*.
- Catalán: se ha partido del WN 1.6 (libre para esta lengua) y se ha construido una versión preliminar del WN 3.0 mediante un *mapping*. El resultado se ha revisado y ampliado manualmente. En el momento de realizar este trabajo esta versión de WN 3.0 contaba con 41.079 *synsets* y 68.876 *variants*.

4.1. Evaluación del algoritmo A

En la figura 2 podemos observar los resultados obtenidos mediante el algoritmo A para el inglés (la lengua original) y el español y catalán (utilizando Google Translate). Es muy importante recordar que el inglés es la lengua original y que por este motivo los valores obtenidos son los mejores. Estos valores nos pueden dar una idea de los resultados que podríamos obtener con un traductor automático perfecto. El español y el catalán han utilizado el mismo traductor automático y sin embargo los resultados para el castellano son notablemente mejores. Esto es probablemente debido a la mejor calidad del traductor de Google para el castellano, ya que probablemente los modelos de lengua y de traducción con los que trabajan, al ser obtenidos principalmente de la web, sean mayores. En estos gráficos en el eje x tenemos el número de *synsets* obtenidos y en el eje y la precisión. Así, por ejemplo, en la figura 2 podemos ver que para el inglés podemos obtener unos 30000 *synsets* con una precisión del 60 %,

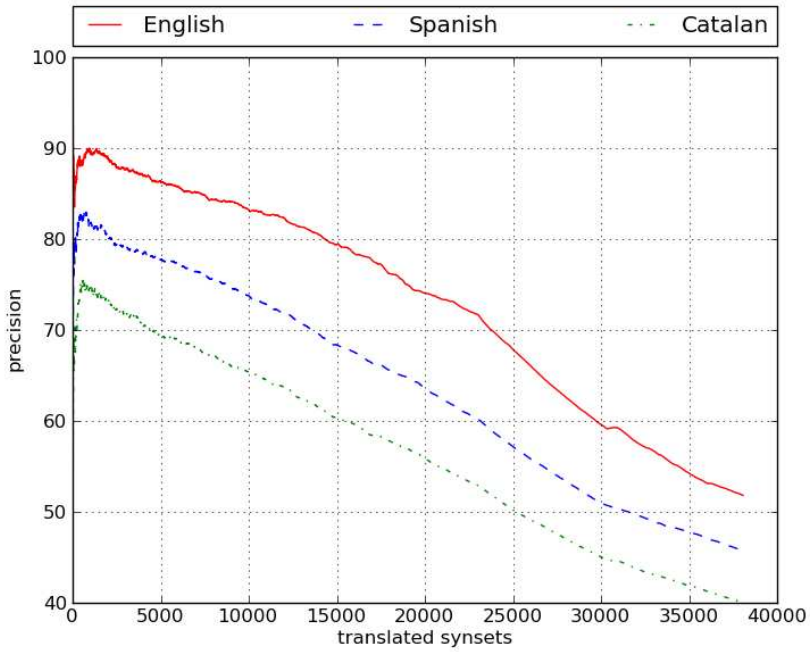

Figura 2: Comparación de los resultados del algoritmo A para el inglés, castellano (Google) y catalán (Google).

Dado que para el inglés obtenemos 14320 *synsets* con una precisión superior al 80 %, si comparamos estos resultados con los presentados en (Atserias et al., 1997) y (Benítez et al., 1998) vemos que tanto para el castellano como para el catalán obtenemos resultados comparables a los de las palabras polisémicas de los grupos 3 y 4 (y para el castellano incluso del grupo 1 y 2). Así, para el catalán, si consideramos la precisión obtenida para el grupo 3 de polisémicas (71.7 %), mediante el algoritmo A obtenemos 3.156 *synsets*. Si consideramos la precisión para el grupo 4 de polisémicas (54.5 %) obtenemos 21.270 *synsets*. Para el castellano, como ya hemos comentado, los resultados son mejores. En este caso, además de obtener resultados comparables para los grupos 3 y 4 de polisémicas, obtenemos también 2160 synstes con la precisión de grupo 1 de polisémicas (80 %) y 8774 con la precisión del grupo 2 de polisémicas (75 %).

Si la evaluación de este algoritmo la llevamos a cabo exclusivamente para los sustantivos (recordemos que en (Atserias et al., 1997) y (Benítez et al.,

1998) se desarrollaron únicamente *synsets* para esta categoría gramatical) observamos que para el catalán podemos obtener 6908 *synsets* con una precisión superior al 71.7 % y para el castellano 5244 *synsets* con una precisión superior al 80 %.

En este mismo experimento hemos realizado una comparación para los dos traductores utilizados para el español (Google Translate y Microsoft) y para una combinación de ambos. Esta combinación ha consistido simplemente en tomar los resultados de los dos traductores a la vez. Los valores obtenidos para ambos traductores son prácticamente idénticos y su combinación aumenta la precisión muy ligeramente. En la figura 3 podemos observar los valores obtenidos.

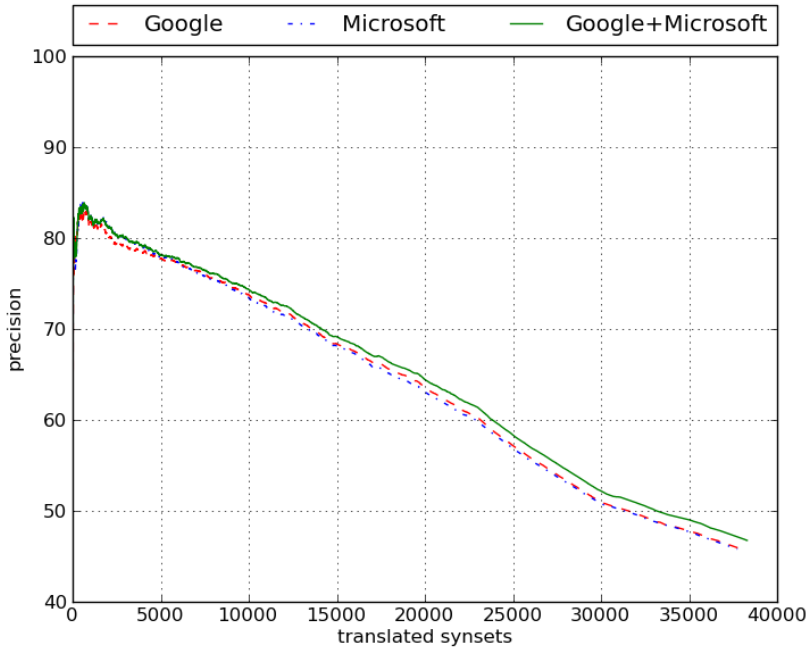

Figura 3: Resultados correspondientes al algoritmo A para el castellano para los dos traductores y la combinación.

4.2. Evaluación del algoritmo B

En la figura 4 podemos observar los resultados obtenidos mediante el algoritmo B para el inglés, español y catalán. Es importante notar que la escala

del eje x correspondiente a los *synsets* traducidos ha cambiado respecto las figuras anteriores. Mediante el algoritmo B hemos podido mejorar notablemente la precisión a costa de reducir el número de *synsets* para los que se han obtenido *variants*.

Para el inglés podemos obtener 10876 *synsets* con una precisión superior al 90 % y 15241 con una precisión superior al 80 %. Para el catalán obtenemos 1029 *synsets* con una precisión comparable al grupo 1 de las polisémicas (90.4 %) y 9938 al grupo 2 de las polisémicas (77.9 %). Para el español los resultados son comparables a los grupos 1, 2 y 3 de las monosémicas por dos motivos, por un lado porque los resultados del algoritmo B para el español son mejores, y por otro lado porque los resultados de los artículos citados son peores para el castellano. Así, obtenemos 1527 *synsets* con una precisión comparable al grupo 1 de las monosémicas (92 %) y 7149 comparables a los grupos 2 y 3 de las monosémicas (89 %).

Si evaluamos los resultados aislados para los sustantivos observamos que para el catalán podemos obtener 1014 *synsets* con una precisión superior al 90.4 % y para el castellano 1003 *synsets* con una precisión superior al 92 %.

En la figura 5 podemos observar la comparación de los resultados correspondientes al algoritmo A y B para el catalán, respetando la escala del eje x de la figura 4. Se puede observar que la precisión aumenta aproximadamente unos 15 puntos.

5. Conclusiones y trabajo futuro

En este artículo hemos presentado una metodología de construcción de WordNets basada en traducción automática de corpus etiquetados semánticamente. Los resultados son comparables a los obtenidos para las mismas lenguas mediante uso de diccionarios bilingües. En global, si se toman todos los resultados presentados para el castellano (Atserias et al., 1997) que obtienen una precisión superior al 85 % se obtienen 7.131 *synsets* con una precisión global del 87.4 %. Nuestro algoritmo B es capaz de obtener 8.098 synset en las mismas condiciones. Posteriormente, (Atserias et al., 1997) combinan los resultados de los métodos descartados y consiguen obtener 10.786

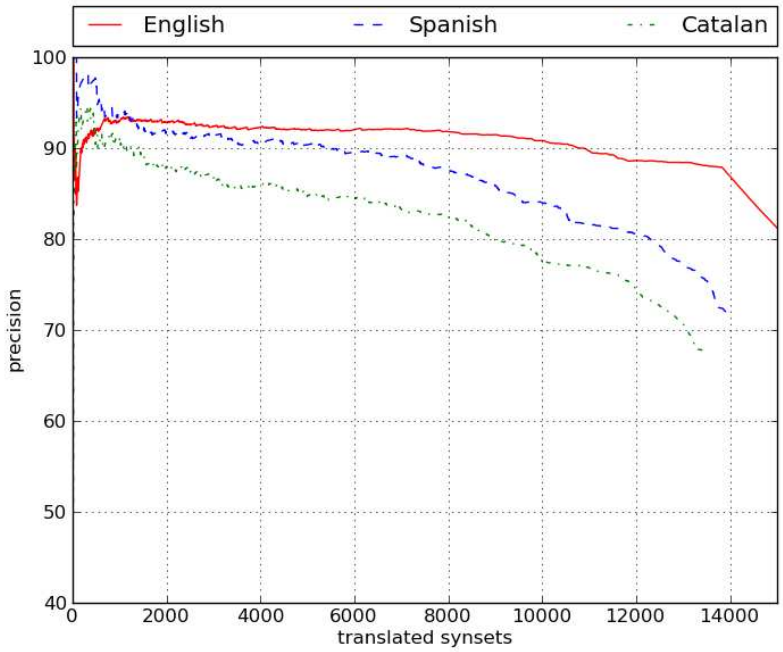

Figura 4: Resultados correspondientes al algoritmo B para el inglés, español y catalán.

con una precisión del 86.4 % mientras que nuestro algoritmo B es capaz de obtener 8643 *synsets*. Si realizamos la misma comparación para el catalán con los resultados presentados en (Benítez et al., 1998) observamos que obtienen unos resultados superiores que para el castellano, mientras que nuestra propuesta funciona peor para el catalán. Mientras que tomando todos los resultados de los métodos basados en diccionario que obtienen una precisión superior al 85 % se obtienen 7.050 *synsets* con una precisión global del 93.8 %, nuestro algoritmo B no obtiene *synsets* con tal precisión. Esta comparación nos sugiere que aunque los resultados son comparables, es necesario buscar estrategias de mejora. También nos muestra que el método es muy sensible a la calidad del traductor automático.

Cabe destacar que la evaluación de los algoritmos se ha llevado a cabo de manera automática a partir de una versión preliminar del WordNet 3.0 para el castellano y catalán no completa. Por este motivo, pueden haber propuestas correctas de los algoritmos que no hayan sido evaluadas como

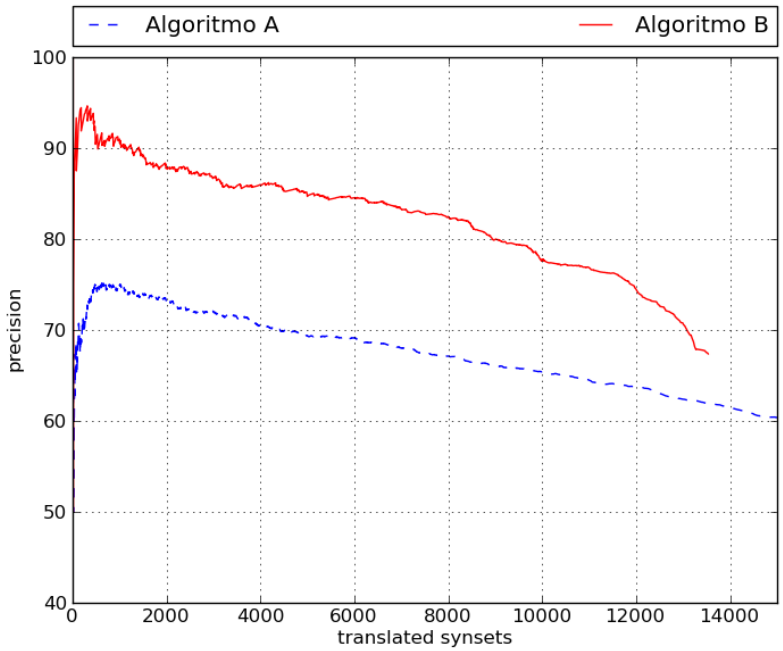

Figura 5: Comparación de los algoritmos A y B para el catalán.

tales, por lo que la precisión real puede ser algo mayor.

La ventaja de la metodología que proponemos es que se puede aplicar a cualquier lengua que disponga de un traductor automático capaz de desambiguar sentidos (ya sea un traductor estadístico o de otro tipo que incorpore desambiguación por sentidos)[9] y un etiquetador morfonsintáctico. Los algoritmos presentados se pueden mejorar con diversas estrategias: uso de la información ya extraída, uso de diccionarios bilingües auxiliares, etc. También se explorará el uso de alineadores estadísticos, como GIZA++ (Och y Ney, 2003) o el alineador de Berkeley (Liang, Taskar, y Klein, 2006). En trabajos futuros abordaremos estas posibles mejoras.

Queda también para un trabajo futuro la solución de las dos limitaciones comentadas: la obtención de *variants* multipalabra y la obtención de más de una *variants* por *synset* cuando sea pertinente.

Otro problema que se debe afrontar en un trabajo futuro es la ampliación del límite máximo de cobertura que viene dado por los *synsets* presentes

[9]Google Translate ofrece traducción del inglés a más de 50 lenguas.

en los corpus utilizados. En este sentido puede ser interesante la aproximación propuesta en el proyecto BabelNet, donde se consideran los enlaces entre páginas de la Wikipedia como un pseudo-etiquetado semántico. De esta manera podemos obtener un corpus que contenga *synsets* no presentes en el corpus de partida. Otra vía de estudio para aumentar el corpus es el uso de etiquetadores semánticos automáticos (Padró et al., 2010b).

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.

Azarova, I., O. Mitrofanova, A. Sinopalnikova, M. Yavorskaya, y I. Oparin. 2002. Russnet: Building a lexical database for the russian language. En *Workshop on WordNet Structures and Standarisation, and how these affect WordNet Application and Evaluation*, páginas 60–64, Las Palmas de Gran Canaria (Spain).

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*.

Cilibrasi, R. L y P. M.B Vitanyi. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.

Dagan, I. y A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

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

Fernández-Montraveta, A., G. Vázquez, y C. Fellbaum. 2008. The Spanish Version of WordNet 3.0. *Text resources and Lexical Knowledge*, páginas 175–182.

Liang, P., B. Taskar, y D. Klein. 2006. Alignment by agreement. En *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, páginas 104–111. Association for Computational Linguistics.

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.

Och, F.J. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Padró, L., M. Collado, S. Reese, M. Lloberes, y I. Castellón. 2010a. FreeLing 2.1: Five years of open-source language processing tools. En *LREC*, volumen 10, página 931–936.

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

Saveski, M. y I. Trajkovski. 2010. Automatic construction of wordnets by using machine translation and language modeling. En *13th Multiconference Information Society*, Ljubljana, Slovenia.

Tufis, D., D. Cristea, y S. Stamou. 2004. BalkaNet: aims, methods, results and perspectives. a general overview. *Science and Technology*, 7(1-2):9–43.

Vossen, P. 1996. Right or wrong. combining lexical resources in the EuroWordNet project. En *Proceedings of Euralex-96*, página 715–728, Goetheborg.

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