WN-Toolkit: un toolkit per a la creació de WordNets a partir de diccionaris bilingües*

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Resum

En aquest article presentem un conjunt de programes que faciliten la creació de WordNets a partir de diccionaris bilingües mitjançant l'estratègia d'expansió. Els programes estan escrits en Python i són per tant multiplataforma. El seu ús, tot i que no disposen d'una interfície gràfica d'usuari, és molt senzill. Aquests programes s'han fet servir amb èxit en el projecte Know2 per a la creació de les versions 3.0 dels WordNets català i espanyol. Els programes estan publicats sota una llicència GNU-GPL i es poden descarregar lliurement de http://lpg.uoc.edu/wn-toolkit.

Paraules clau

WordNet, estratègia d'expansió

Abstract

This paper presents a set of programs to facilitate the creation of WordNet from bilingual dictionaries following the expand model. The programs are written in Python and are therefore multiplatform. The programs are very easy to use although they don't have a graphical user interface. These programs have been successfully used in the Know2 Project for the creation of Catalan and Spanish WordNet 3.0. The programs are published under the GNU-GPL licence and can be freely downloaded from http://lpg.uoc.edu/wn-toolkit.

Keywords

WordNet, expand model

1 Introducció

WordNet (Fellbaum, 1998) és una base de dades lèxica de l'anglès desenvolupada a la Universitat de Princeton (per aquest motiu, a la resta de l'article anomenarem a aquesta versió PWN - *Princeton WordNet*). En aquesta base de dades les paraules que pertanyen a categories obertes (és a dir, els substantius, verbs, adjectius i adverbis) s'organitzen en conjunts de sinònims denominats *synsets*. Cada *synset* representa un concepte lexicalitzat en anglès, i es connecta amb els altres *synsets* mitjançant una sèrie de relacions semàntiques: hiponímia o relació d'especificitat entre un mot (l'hipònim) i un altre de significat més genèric (hiperònim), l'antonímia o relació entre mots que tenen un significat directament oposat, la meronímia o la relació entre una part i un tot, i la troponímia o implicació lèxica, una relació que es dóna entre verbs i que es pot considerar en certa manera equivalent a la relació d'hiponímia per als substantius. Per exemple, el *synset* del WordNet 3.0 de l'anglès que s'identifica mitjançant un *offset* i una categoria gramatical 02958343-n té assignades diverses variants: *car, auto, automobile, machine i motorcar*. Cada *synset* té assignada una glossa o definició en alguns casos també exemples d'ús. Per al nostre exemple *a motor vehicle with four wheels; usually propelled by an internal combustion engine. he needs a car to get to work.* Aquest *synset* té 31 hipònims, com per exemple 02701002-n (*ambulance*) o 03594945-n (*jeep, landrover*), entre d'altres. També té un hiperònim, 03791235-n (*motor vehicle, automobile vehicle*). D'entre els merònims registrats a WordNet podem posar com a exemple 02685365-n (*airbag*).

1.1 El PWN

El WordNet anglès és lliure i es pot descarregar de la plana web de la Universitat de Princeton[1]. La versió actual és la 3.1 però aquesta versió per ara només es pot consultar *online*. La versió 3.0 es pot descarregar i en aquest article ens centrarem en aquesta versión. A la taula 1 podem observar una comparativa sobre el nombre de *syn-*

*Aquest treball s'ha portat a terme dins del projecte Know2 *Language understanding technologies for multilingual domain-oriented information access* (MICINN, TINN2009-14715-C04-04)

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

sets per a les versions 1.5, 1.6 y 3.0 del PWN. Com podem veure, el nombre de *synsets* desenvolupats augmenta amb les noves versions.

	1.5	**1.6**	**3.0**
Total	76.705	99.642	118.695
Substantius	51.253	66.025	83.073
Verbs	8.847	12.127	13.845
Adjectius	13.460	17.915	18.156
Adverbis	3.145	3.375	3.621

Taula 1: Comparació del nombre de *synsets* per a tres versions del PWN

1.2 Wordnets per a altres llengües

En diversos projectes s'han desenvolupat wordnets per a altres llengües: EuroWordNet (Vossen, 1998) inicialment per a l'holandès, italià, espanyol i l'ampliació i millora de l'anglès, i en una extensió del projecte per a l'alemany, francès, estonià i txec; el projecte Balkanet (Tufis, Cristea i Stamou, 2004) per al búlgar, grec, romanès, serbi i turc i el projecte RusNet (Azarova et al., 2002) per al rus, entre d'altres projectes similars. A la plana web de la *Global WordNet Association*[2] podem trobar una llista dels WordNets disponibles per a les diferents llengües i amb el seu estat de desenvolupament.

No tots els WordNets que es construeixen es publiquen amb una llicència lliure. A Bond i Kyonghee (2008) podem trobar els WordNets existents per a les diferents llengües amb la llicència associada. Les llengües que disposen de WordNets lliures són: anglès, finès, rus, tailandès, danès, japonès, català, gaèlic, hindi, francès, malai, indonesi, castellà, gallec, basc, àrab i hebreu. De la plana web del projecte *Open Multilingual WordNet* (Bond i Paik, 2012) es poden descarregar un gran nombre de WordNets lliures en un format unificat.

El projecte que ha permès desenvolupar el *toolkit* que presentem en aquest article té com a objectiu desenvolupar els WordNets 3.0 per al català i castellà i distribuir-los sota una llicència lliure.

1.2.1 Estratègies per a la construcció de WordNets

Podem agrupar les estratègies generals per a la construcció de wordnets en dos gran grups: (Vossen, 1998):

- **Estrategia de combinació** (*merge model*): per a cada llengua es genera una ontologia amb els seus propis nivells i relacions. Posteriorment es generen relacions interlingüístiques entre aquesta ontologia i el PWN.
- **Estrategia d'expansió** (*expand model*): es tradueixen les *variants* associades als *synsets* del PWN, fent servir diferents estratègies. Després es verifica si les relacions entre *synsets* donades pel PWN són també vàlides per a la llengua d'arribada.

Vossen (1996) enumera una sèrie d'avantatges i inconvenients per a cada una d'aquestes estratègies. L'estratègia d'expansió és tècnicament més senzilla i garanteix un grau més alt de compatibilitat entre els WordNets de les diferents llengües. Però els WordNets desenvolupats d'aquesta manera estan molt influenciats pel PWN i contindran tots els seus errors i deficiències estructurals. L'estratègia de combinació és més complexa però permet un major aprofitament més directe de les ontologies i tesaurus disponibles.

1.3 Estratègia d'expansió

Com hem comentat, l'estratègia d'expansió consisteix a construir un WordNet per a una llengua determinada traduint les *variants* assignades a cada *synset* del PWN. La manera més evident i emprada és mitjançant l'ús de diccionaris bilingües. La principal dificultat per aplicar l'estratègia d'expansió és la polisèmia. Si totes les *variants* fossin monosèmiques, és a dir, que estiguéssin assignades a un únic *synset* el problema seria simple, ja que només caldria trobar una o mes traduccions per a la *variant* anglesa. Com que la paraula anglesa només tindria un sentit la traducció d'aquesta paraula seria la *variant* correcta en la llengua d'arribada.

Veiem aquest fet amb un exemple. La paraula *aeroplane* està assignada a un únic *synset* (02691156-n) i és, per tant, monosèmica segons el PWN. Si consultem un diccionari anglès-català trobarem dues possibles traduccions: avió i aeroplà. Aquestes dues paraules seran *variants* vàlides per a aquest *synset*. Per una altra banda, la paraula *plane* és polisèmica segons el PWN, ja que està assignada a més d'un *synset* (de fet està assignada als següents *synsets*: 02691156-n, 13861050-n, 13941806-n i 03954731-n). Si traduïm *plane* amb un diccionari anglès-català trobarem les següents possibles traduccions: avió, pla, nivell, garlopa, ribot, plàtan. En aquest cas

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

no podem assignar totes aquestes paraules com a *variants* vàlides per a tots els *synsets* ni tampoc disposem de suficient informació per a saber a quin *synset* assignar cada una de les *variants*.

A la taula 2 podem observar el nombre de *variants* que tenen assignades un nombre determinat de *synsets*. Les *variants* que tenen assignat un únic *synset* són paraules monosèmiques en anglès (almenys segons PWN). Així, per exemple, el 82.32% de les *variants* del PWN són monosèmiques.

N. 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

Taula 2: Nombre de *variants* que tenen assignades un nombre determinat de *synsets*

Un altre aspecte interessant és observar si aquestes *variants* estan escrites amb la primera lletra en majúscula (i correspondran probablement a un nom propi) i quantes estan escrites amb totes les lletres en minúscules. A la taula 3 podem observar aquests valors.

	variants	%
minúscula	84.714	68.75
majúscula	38.514	31.25

Taula 3: Nombre de *variants* monosèmiques del PWN segons estiguin escrites en minúscules o amb la primera lletra en majúscula

2 Els WordNets per al català i castellà

Els WordNets del castellà (Atserias et al., 1997) i del català (Benítez et al., 1998) es van construir seguint una metodologia d'expansió, ja que es van traduir les *variants* corresponents als *synsets* del PWN. Per als substantius es va fer servir una metodologia basada en diccionaris bilingües. En canvi, els verbs es van desenvolupar d'una manera manual i els adjectius i adverbis no es van desenvolupar en les primeres versions.

A la taula 4 podem observar el nombre de synsets per a les versions 1.6 i 3.0 dels WordNets de l'anglès, català i castellà[3].

1.6	Anglès	Català	Castellà
Total	99.645	41.991	21.252
N	66.025	32.236	11.218
V	12.127	5.397	4.994
A	17.915	4.358	5.040
R	3.375	0	0
3.0	**Anglès**	**Català**	**Castellà**
Total	118.695	46.033	38.702
N	83.073	36.460	26.594
V	13.845	5.424	6.251
A	18.156	4.148	5.180
R	3.621	1	677

Taula 4: Nombre de *variants* per a les versions 1.6 i 3.0 de l'anglès, català o castellà

3 Organització de l'article

En aquest article presentem tant el propi *toolkit* com una avaluació dels mètodes que implementa per a la creació dels WordNets 3.0 per al català i castellà. Aquest *toolkit* està format pels programes que hem fet servir en la nostra recerca més una petita documentació per a cada un dels programes. L'objectiu és posar a disposició de la comunitat aquests programes amb l'esperança que sigui útils per a la creació de WordNets per a altres llengües.

La resta de l'article està organitzat de la següent manera. En primer lloc presentarem la tecnologia emprada per a la creació dels programes i els requisits necessaris per a poder-los executar, així com a les instruccions que són comunes per a tots aquests programes. Posteriorment presentarem algunes de les estratègies emprades per a la construcció dels WordNets 3.0 del català i castellà, que són:

- Ús de diccionaris bilingües
- Ús de Babelnet

En la construcció dels WordNets 3.0 per al català i castellà també s'ha fet servir una estratègia basada en l'explotació de corpus paral·lels (Oliver i Climent, 2011; Oliver i Climent, 2012a; Oliver i Climent, 2012b). En aquesta versió del *toolkit* no estan presents els programes necessaris per fer servir aquesta estratègia. Aquests programes es distribuiran en futures versions del *toolkit*.

Per a cada una d'aquestes estratègies es presentarà:

- Una descripció detallada de l'estratègia
- Els programes necessaris per construir WordNets amb aquesta estratègia

[3]La versió 1.6 del castellà no és lliure i en aquesta taula mostrem els valors del fragment lliure distribuït amb l'analitzador Freeling

- Els resultats de l'avaluació d'aquesta estratègia en la construcció dels WordNets 3.0 per al català i castellà

S'ha dut a terme una avaluació automàtica, consistent en comparar els resultats obtinguts amb les versions preliminars dels WordNets 3.0 del català i castellà. Si una *variant* obtinguda per a un determinat *synset* coincideix amb alguna de les presents en les versions preliminars, aquesta es dóna per correcta. Si no coincideix amb cap de les presents, es considera incorrecta. En cas de no tenir cap *variant* per al *synset* corresponent en les versions preliminars, no s'avalua aquest resultat. Som conscients que els resultats d'aquesta avaluació automàtica poden variar considerablement dels d'una avaluació manual i per aquest motiu en molts casos s'ha portat a terme també una avaluació manual.

4 Aspectes generals sobre el toolkit

El *toolkit* que presentem en aquest article està format per un conjunt de programes escrits en Python. Python és un llenguatge interpretat i l'únic que necessitem per executar aquests programes és disposar del corresponent intèrpret. L'intèrpret es pot descarregar gratuïtament de `http://www.python.org`. Hi ha versions per als sistemes operatius més habituals. Linux i Mac acostumen a tenir instal·lat l'intèrpret de Python per defecte, de manera que si treballeu amb aquests sistemes operatius no necessitareu instal·lar res al vostre ordinador. Els programes que presentem no disposen d'interfície gràfica d'usuari i funcionen sota línia de comandes (*Terminal* en Linux i Mac i *Símbol de sistema* en Windows).

Els programes s'executen donant una sèrie de paràmetres que variarà segons el programa. Per saber els paràmetres que cal donar podeu fer:

```
python nomprograma.py -h
```

on *nomprograma.py* és el nom del programa que volem executar.

El Toolkit es pot descarregar de `http://lpg.uoc.edu/wn-toolkit`.

5 Ús de diccionaris bilingües

5.1 Descripció de l'estratègia

Amb aquesta primera estratègia obtenim *variants* únicament per als *synsets* les *variants* en anglès dels quals són monosèmiques. És a dir, traduïm mitjançant diferents tipus de diccionaris (generals, enciclopèdics i terminològics) paraules angleses monosèmiques (assignades a un únic *synset*) i assignem aquest *synset* a la corresponent paraula o paraules de la llengua d'arribada donades pel diccionari.

5.2 Programes

Per fer ús d'aquesta estratègia hem de fer servir diversos programes:

- **createmonosemicwordlist.py**: per crear les llistes de paraules monosèmiques del PWN anglès. Alternativament, podem fer servir directament les llistes de paraules monosèmiques que es distribueixen amb el *toolkit* corresponents a la versió 3.0.
- **wndictionary.py**: a partir d'una llista de paraules monosèmiques del PWN anglès i d'un diccionari bilingüe proporciona una llista de *synsets* amb les seves corresponents *variants* en la llengua d'arribada.
- **apertium2bildic.py**: crea un diccionari bilingüe adequat per fer-lo servir amb el programa wndictionary.py a partir dels diccionaris de transferència del sistema de traducció automàtica Apertium (Forcada, Tyers i Ramírez-Sánchez, 2009).
- **dacco2bildic.py**: crea un diccionari bilingüe adequat per fer-lo servir amb el programa wndictionary.py a partir del diccionari de lliure distribució anglès-català Dacco[4]
- **TO2bildic.py**: crea un diccionari bilingüe adequat per fer-lo servir amb el programa wndictionary.py a partir dels glossaris terminològics Terminologia Oberta del TermCat[5].
- **wiktionary2bildic.py**: crea un diccionari bilingüe adequat per fer-lo servir amb el programa wndictionary.py a partir dels fitxers dump xml del Wiktionary[6]
- **wikipedia2bildic.py**: crea un diccionari enciclopèdic bilingüe adequat per fer-lo servir amb el programa wndictionary.py a partir del fitxers dump xml de la Vikipèdia[7]
- **combinedictionary.py**: aquest programa permet combinar diversos diccionaris, de manera que es crea un únic diccionari que conté la informació de tots ells, sense reduplicar la informació comuna.

[4] `http://www.catalandictionary.org/`
[5] `http://.termcat.org`
[6] `www.wiktionary.org`
[7] `www.wikipedia.org`

En els següents subapartats expliquem amb més detall cada un d'aquests programes.

5.2.1 createmonosemicwordlist.py

Aquest programa extreu tres llistes de *variants* monosèmiques del PWN. Per poder executar el programa primer hem de descarregar els arxius corresponents a les bases de dades del WordNet des de `http://wordnetcode.princeton.edu/3.0/WNdb-3.0.tar.gz` (o l'arxiu corresponent a la versió desitjada). El programa pren com a paràmetres el directori on es troben els arxius de WordNet i el prefix que volem que tinguin els arxius de sortida. Es creen tres arxius: un que conté totes les variants monosèmiques, un que conté les que estan escrites en minúscules i un altre que conté les que estan escrites amb la primera lletra en majúscula. Per executar el programa podem fer:

```
python createmonosemicwordlist.py -d ./dict
-p pwnmonosemic
```

Els arxius de sortida contenen l'offset i categoria gramatical i la variant monosèmica separats per una tabulador, com al següent exemple:

```
02691156n aeroplane
02691156n airplane
```

Amb el *toolkit* es distribueixen els fitxers corresponents a les variants monosèmiques de la versió 3.0.

5.2.2 wndictionary.py

Amb aquest programa podem obtenir las variants en la llengua d'arribada a partir d'una llista de variants angleses monosèmiques obtingudes amb createmonosemicwordlist.py. El programa demana el fitxer de diccionari que volem fer servir, el fitxer d'entrada amb les variants monosèmiques i el nom del fitxer de sortida. Per executar el programa podem fer:

```
python wndictionary.py -d diccionari.txt
-i pwnmonosemic.txt -o wordnetcreat.txt
```

La sortida té la següent forma:

```
02691156n avión
02691156n aeroplano
```

5.2.3 apertium2bildic.py

Aquest programa permet crear un diccionari bilingüe a partir dels diccionaris de transferència del sistema de traducció automàtica Apertium[8]. Aquest sistema de traducció automàtica es distribueix sota llicència lliure i es poden descarregar les dades lingüístiques des de `http://sourceforge.net/projects/apertium/files/`. Triarem el parell de llengües desitjat i descarregarem l'arxiu corresponent a la darrera versió. Un cop descomprimit l'arxiu seleccionarem l'arxiu corresponent al diccionari de transferència (que s'anomena per l'anglès-català: apertium-en-ca.en-ca.dix).

```
python apertium2bildic.py -a
apertium-en-ca.en-ca.dix -o
diccionari-en-ca.txt
```

En els casos en què l'anglès no estigui com a primera llengua del diccionari, caldrà fer servir l'opció -r, per canviar l'ordre de les entrades i tenir l'anglès en primer lloc. La sortida és un fitxer de text tabulat amb paraula en anglès, categoria gramatical i paraula en la llengua d'arribada, com al següent exemple:

```
caution n amonestació
cautious a cautelós
cautiously r amb cautela
```

5.2.4 dacco2bildic.py

Dacco és un diccionari lliure col·laboratiu anglès-català i català-anglès. Es poden descarregar els arxius de `http://sourceforge.net/projects/dacco/`. Descarreguem l'arxiu dacco-0.9.zip (o el corresponent a la darrera versió disponible) i el descomprimim. Podem executar el programa donant com a paràmetres el directori on es troben els diccionaris anglès-català i el fitxer de sortida, per exemple:

```
python dacco2bildic.py
./Dacco-0.9/dictionaries/engcat
-o diccionaridacco.txt
```

La sortida és un fitxer de text tabulat amb paraula en anglès, categoria gramatical i paraula en català, com al següent exemple:

```
last a últim:darrer
last v durar
last name n cognom
```

5.2.5 TO2bildic.py

Aquest programa transforma un o més glossaris terminològics de Terminologia Oberta del TermCat en un diccionari en format per ser utilitzat amb el programa wndictionary.py. Els glossaris de Terminologia Oberta del TermCat es poden descarregar de `http://www.termcat.cat/productes/toberta.htm`. Per fer funcionar aquest programa hem de descarregar almenys un dels glossaris (es pot treballar amb més d'un alhora i fins i tot amb tots). Posem tots els glossaris en un directori i els descomprimim. Al programa passarem com a paràmetres el glossari a tractar o el directori que conté tots els glossaris que volem tractar i el fitxer de sortida. Per exemple, si volem tractar tots els glossaris que estan al directori /home/TO i posar el resultat en un arxiu anomenant TO.txt hem de fer:

[8] `http://www.apertium.org`

```
python TO2bildic.py /home/TO -O TO.txt
```

Donat que els glossaris de Terminologia Oberta no especifiquen la categoria gramatical, assignem "n"(substantiu) a totes les entrades, ja que la immensa majoria de les entrades corresponen a aquesta categoria gramatical. La sortida és un fitxer de text tabulat amb paraula en anglès, categoria gramatical i paraula en la llengua d'arribada, com al següent exemple:

```
fifth disease n eritema infecciós
fight n baralla
fighting n baralla
```

5.2.6 wiktionary2bildic.py

Aquest programa serveix per a generar un diccionari bilingüe a partir dels fitxers *dump xml* de Wiktionary[9]. Wiktionary és un projecte col·laboratiu per a la creació de diccionaris en moltes llengües, amb enllaços interlingüístics. És possible descarregar els fitxers *dump xml* de `http://dumps.wikimedia.org/`. Com que per a la creació de WordNets ens interessen diccionaris d'anglès a la llengua d'arribada, descarregarem els fitxers corresponents al Wiktionary anglès, des de `http://dumps.wikimedia.org/enwiktionary/`. És recomanable fer servir sempre la darrera versió disponible. Cal tenir en compte que aquests fitxers són molt grans i que la descàrrega pot durar molt de temps. També cal preveure tenir espai en disc disponible.

```
python wiktionary2bildic.py
enwiktionary-latest-pages-articles.xml
-l Catalan -o wiktionary-eng-cat.txt
```

Cal tenir en compte que les llengües s'especifiquen amb el nom complet en anglès (Catalan, Spanish, French...) A continuació observem una mostra del resultat:

```
listen v escoltar
literally r literalment
literary a literari
```

5.2.7 wikipedia2bildic.py

Aquest programa és semblant al wiktionary2bildic.py però crea un diccionari a partir dels fitxers *dump xml* de Wikipedia[10]. Els diccionaris que es creïn seran uns diccionaris de caire enciclopèdic donada la pròpia naturalesa de la Vikipèdia. És possible descarregar els fitxers *dump xml* de `http://dumps.wikimedia.org/`. Com que per a la creació de WordNets ens interessen diccionaris d'anglès a la llengua d'arribada, descarregarem els fitxers corresponents a la Vikipèdia anglesa, des de `http://dumps.wikimedia.org/enwiki/`. És recomanable fer servir sempre la darrera versió disponible. Cal tenir en compte que aquests fitxers són molt grans i que la descàrrega pot durar molt de temps. També cal preveure tenir espai en disc disponible.

Per executar el programa simplement s'ha de fer:

```
python wikipedia2bildic.py
enwiki-latest-pages-articles.xml
-l ca -o wikipedia-eng-cat.txt
```

Cal tenir en compte que les llengües s'especifiquen amb els codi ISO de dues lletres (ca, es, fr...) A continuació observem una mostra del resultat:

```
Gregor Mendel n Gregor Mendel
Grammar n Gramàtica
Gigabyte n Gigabyte
Galaxy groups and clusters n
Cúmul de galàxies
```

A la Vikipèdia no disposem d'informació sobre la categoria gramatical, però la immensa majoria seran substantius i per aquest motiu assignem la categoria n a totes les entrades.

5.2.8 combinedictionary.py

Aquest programa permet combinar dos o més diccionaris en un de sol. Té cura de no repetir ni entrades ni accepcions. Els paràmetres que cal donar són dos o més diccionaris d'entrada seguit del nom del diccionari de sortida. El programa verifica que el fitxer de sortida no existeixi, per evitar sobreescriure un fitxer existent.

```
python combinedictionary.py dict1.txt
dict2.txt dict3.txt dictsortida.txt
```

5.3 Avaluació

5.3.1 Ús de diccionaris bilingües generals

En aquest experiment hem generat els WordNets a partir de diccionaris obtinguts a partir dels diccionaris de transferència d'Apertium i del Wiktionary. A la taula 5 podem observar el nombre d'entrades de cada un d'aquests diccionaris, així com el nombre d'entrades del diccionari resultant de combinar les dues fonts.

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

Taula 5: Nombre d'entrades dels diccionaris bilingües generals

Per al castellà podem obtenir un total de 12.676 *variants* de las que 7.401 són correctes, 2.997 incorrectes (segons l'avaluació automàtica) i no podem avaluar 2.278. La precisió per al castellà, segons l'avaluació

[9] `http://www.wiktionary.org`
[10] `http://www.wikipedia.org`

automàtica és del 71.2%. S'han revisat manualment tots els resultats considerats incorrectes per l'avaluació automàtica. Això ens ha permès calcular una nova precisió, que ara puja fins el 93.95%.

Per al català obtenim un total de 8.335 *variants*, de les que 4.223 són correctes, 1.083 incorrectes (segons l'avaluació automàtica) y no podem avaluar 3.029. La precisió per al català, segons l'avaluació automàtica, és del 79.6%. De la mateixa manera que per al castellà, hem revisat els resultats incorrectes i hem pogut calcular una nova precisió que ara puja fins el 96.36%.

5.3.2 Ús de diccionaris enciclopèdics

En aquest experiment s'ha fet servir un diccionari enciclopèdic per a traduir les *variants* angleses monosèmiques escrites amb la primera lletra en majúscula. Aquestes constitueixen el 31.15% de les *variants* monosèmiques, com es pot veure a la taula 3. D'aquestes, la immensa majoria (99.17%) són substantius.

S'ha creat un diccionari enciclopèdic bilingüe anglès castellà de 59.659 entrades i anglès-català de 22.205 entrades a partir de la Vikipèdia anglesa fent servir el programa wikipedia2bildic.py.

Per al castellà podem obtenir un total de 10.356 *variants* de les que 4.722 són correctes. 1.916 incorrectes (segons l'avaluació automàtica) i no podem avaluar 3.718. La precisió per al castellà, segons aquesta avaluació automàtica, és del 71.1%. Si revisem els casos donats per incorrectes i recalculem la precisió, aquesta augmenta fins el 89.74%.

Per al català obtenim un total de 7.083 *variants*, de les que 2.642 són correctes, 1.278 incorrectes (segons l'avaluació automàtica) i no podem avaluar 3.163. La precisió per al català, segons l'avaluació automàtica, és del 67.4%. Després de la revisió manual dels classificats com a incorrectes, la precisió augmenta fins el 90.94%.

5.3.3 Ús de diccionaris terminològics

En aquest experiment hem fet servir un conjunt de diccionaris terminològics per a traduir les *variants* angleses monosèmiques, tant les que estan escrites en minúscules com les que tenen la primera lletra en majúscula. Hem obtingut un diccionari terminològic fent servir el programa TO2bildic.py a partir de tots els glossaris terminològics de Terminologia Oberta del TermCat. D'aquesta manera hem confeccionat un diccionari terminològic anglès-castellà de 46.761 entrades i un anglès-català de 46.653 entrades.

Per al castellà obtenim un total de 10.456 *variants*, de les que 4.180 són correctes, 3.346 incorrectes (segons l'avaluació automàtica) i no podem avaluar 2.930. La precisió per al castellà, segons l'avaluació automàtica, és del 55.5%. Aquest resultat és molt baix i decidim revisar manualment tant les avaluades automàticament com a incorrectes, com les no avaluades. Moltes d'aquestes eren en realitat correctes i la nova precisió augmenta fins el 98.57%.

Per al català podem obtenir un total de 9.890 *variants* de les que 3.007 són correctes, 2.614 incorrectes (segons l'avaluació automàtica) i no podem avaluar 4.269. La precisió per al català calculada de manera automàtica és del 53.6% però si revisem manualment els resultats la precisió augmenta dins el 98.36%.

6 Babelnet

6.1 Descripció

Babelnet (Navigli i Ponzetto, 2010) és una xarxa semàntica de grans dimensions que s'ha creat combinant el coneixement lexicogràfic de WordNet amb el coneixement enciclopèdic de la Vikipèdia. D'aquesta manera Babelnet ofereix una relació entre els *synsets* de WordNet i les entrades de la Vikipèdia. Aquesta relació s'ha obtingut tant per a entrades amb títols monosèmics com polisèmics. Així, aprofitant aquest recurs podrem obtenir *variants* en altres llengües independentment si la *variant* anglesa associada al *synset* és monosèmica com si és polisèmica.

Per poder relacionar les dues fonts els autors prenen de WordNet tots els possibles sentits d'una determinada paraula i totes les relacions semàntiques dels *synsets*. De la Vikipèdia prenen totes les entrades i les relacions donades pels enllaços d'hipertext de les pàgines. Aquestes relacions poden ser de diferents tipus i no estan especificades des del punt de vista semàntic. Per establir un *mapping* entre els dos recursos fan servir els anomenats *contextos de desambiguació*. Aquests contextos, per als articles de la Vikipèdia estan formats per les etiquetes de sentit que tenen algunes entrades, els enllaços d'hipertext i les categories. En el cas de WordNet aquests contextos estan formats per tots els sinònims, hiperònims i hipònims, els lemes de les categories obertes de la glossa o definició i les *variants* associades als *synsets* germans, és a dir, els que tenen un hiperònim directe comú. Per establir els *mappings* apliquen els següents criteris:

- Per a totes les planes de la Vikipèdia que tinguin un títol monosèmic tant per la Vikipèdia com per WordNet, s'enllaça directament la plana amb el *synset*.
- Per a la resta de planes es calcula la intersecció dels contextos de desambigüació per a tots els sentits de la Vikipèdia i WordNet.

Aprofitant els enllaços interlingüístics de la Vikipèdia, aquesta relació es pot establir per a totes les llengües que disposin de l'entrada corresponent.

La versió 3.0 de Babelnet oferia un arxiu anomentat babel-to-wordnet-3.0.txt que tenia el següent aspecte:

```
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
```

és a dir, relacionava el títol d'una entrada de la Vikipèdia anglesa amb una *variant* i un *synset* del

Princeton WordNet 3.0. Els experiments que hem portat a terme s'han fet amb aquesta versió i els resultats que oferim a l'apartat d'Avaluació són els obtinguts amb aquesta versió.

La versió actual disponible difereix en el format i el continguts (Navigli i Ponzetto, 2012). La distribució inclou els següents arxius:

- BabelNet API: és una API escrita en Java per accedir a la informació de Babelnet.
- BabelNet precompiled index: es tracta dels índexs precompil·lats
- BabelNet glosses: Aquest és el que farem servir per extreure informació, ja que conté la relació entre els *synsets* de BabelNet i WordNet i les entrades de la Vikipèdia.

Mirem més a fons el contingut del fitxer BabelNet glosses:

```
bn:00001439n
CA WIKI Almirall Almiral l'és el grau militar,
o part del nom del ranc , amb que es coneixen
els caps d'una flota o marina de guerra.
ES WIKI Almirante Almirante es un grado
militar de la marina de guerra que equivale
al de general en otros cuerpos del ejército.
IT WIKI Ammiraglio Il grado di Ammiraglio
è più alto nella gerarchia delle odierne
marine militari .
DE WIKI Admiral Admiral ist der höchste
militärische Dienstgrad in der Marine,
entsprechend dem General des Heeres und der
Luftwaffe.
EN WIKIWN 09771204n the supreme commander
of a fleet; ranks above a vice admiral and
below a fleet admiral
EN WIKI Admiral  Admiral is the rank , or
part of the name of the ranks , of the
highest naval officers.
```

Si agafem la informació de la línia EN WIKIWN, que és el *synset* de WordNet (09771204n), podem saber directament que una possible *variant* en català és *Almirall*. Aquesta informació és directament deduïble des d'aquest fitxer, ja que inclou el català. Si estem construint un WordNet per a una altra llengua, podríem consultar els enllaços interlingüístics de la Vikipèdia anglesa corresponent a l'entrada *Admiral*. Per exemple, podríem deduir que en holandès és *Admiraal*.

El recurs es pot descarregar de la plana web `http://lcl.uniroma1.it/babelnet/` i també ofereix una interfície de consulta.

6.2 Programes

El programa babel2wordnet.py pren com a paràmetres l'arxiu de glosses de Babelnet i, de manera opcional, un diccionari creat per a la llengua d'arribada desitjada mitjançant el programa wikipedia2bildic.py. Si volem generar un WordNet per alguna de les llengües incloses al fitxer de glosses de Babelnet no serà imprescindible indicar un diccionari; si no en donem cap, simplement extraurà la informació continguda al BabelNet. Si proporcionem un diccionari completarà la informació inclosa en el BabelNet. Per a llengües no incloses a Babelnet, es del tot imprescindible proporcionar un diccionari.

El programa, doncs, funciona de la següent manera:

```
python babel2wordnet.py babel-glosses
-l ru -d diccionari_wikipedia-rus.txt
-o babelwordnet-rus.txt
```

El programa també pot intentar unificar les majúscules i minúscules a partir de la informació continguda en el propi WordNet. Per aconseguir això, s'ha de donar el directori on es troben els arxius de WordNet mitjançant el paràmetre -w.

```
python babel2wordnet.py babel-glosses
-l ru -d diccionari_wikipedia-rus.txt
-o babelwordnet-rus.txt
-w /home/usuari/WordNet30
```

Aquesta opció s'ha de fer servir únicament per a aquelles llengües on la capitalització o no de les paraules segueixi el mateix patró que l'anglès, és a dir, noms propis escrits en majúscules.

La sortida que ens proporcionarà serà com la següent:

```
09771204n almirall
08784104n Eólida
03745571n menhir
12154426n Pandanàcia
12960211n Ophioglossum
00149895n soldadura per punts
```

6.3 Avaluació

Aquesta avaluació s'ha dut a terme amb la versió antiga de BabelNet, és a dir, fent servir el fitxer babelto-wordnet-3.0.txt.

Per al castellà obtenim un total de 26.209 *variants*, de las que 14.614 són correctes, 5.065 incorrectes (segons l'avaluació automàtica) i no podem avaluar automàticament 6.530. La precisió per al castellà, segons aquesta avaluació automàtica, és del 74.3%. Revisem manualment tant les avaluades automàticament com a incorrectes, com les no avaluades. Un cop portada a terme aquesta avaluació manual podem calcular un nou valor de precisió, que és ara del 81.02%.

Per al català podem obtenir un total de 18.366 *variants* de les que 9.044 són correctes, 3.548 incorrectes (segons l'avaluació automàtica) i no podem avaluar 5.774. La precisió per al català, segons l'avaluació automàtica, és del 61%. Un cop revisades tant les avaluades automàticament com a incorrectes com les no avaluades podem calcular una nova precisió que és ara del 80.91%.

7 Conclusions

En aquest article hem presentat el WN-Toolkit per a la creació de WordNets seguint l'estratègia d'expansió mitjançant l'ús de diccionaris bilingües. Hem afegit també programes per crear WordNets a partir de Babelnet. Tots aquests algorismes s'han fet servir amb èxit per a la creació dels WordNet 3.0 del català i castellà.

A l'article presentem també els resultats de l'avaluació d'aquesta metodologia. Les estratègies basades en diccionaris només poden obtenir *variants* per a *synsets* que tinguin assignades *variants* monosèmiques. La metodologia basada en Babelnet no presenta aquesta restricció

Per a la creació dels WordNets del català i castellà també es van fer servir mètodes basats en corpus paral·lels(Oliver i Climent, 2011; Oliver i Climent, 2012a; Oliver i Climent, 2012b). Aquestes metodologies basades en corpus paral·lels no presenten la restricció de les metodologies basades en diccionaris pel que fa a la monosèmia de les *variants* angleses assignades als *synsets*. En una propera versió d'aquest Toolkit afegirem els programes necessaris per replicar aquesta metodologia.

Bibliografía

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

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

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

Bond, F. i P. Kyonghee. 2008. A survey of wordnets and their licenses. Em *Proceedings of the 6th International Global WordNet Conference, Matsue (Japan)*, pp. 64–71.

Bond, F. i K. Paik. 2012. A survey of WordNets and their licenses. Em *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, volume 8, pp. 5, Matsue (Japan).

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

Forcada, M. L, F. M Tyers, i G. Ramírez-Sánchez. 2009. The apertium machine translation platform: five years on. Em *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pp. 3–10.

Navigli, R. i S. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, Elsevier*.

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

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

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

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

Tufis, D., D. Cristea, i 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. Em *Proceedings of Euralex-96*, pp. 715–728, Goetheborg.

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