Applying Automatic Methods for the Construction of Multilingual WordNets

L. Benítez
G. Escudero
J. Farreres
G. Rigau

Report LSI–98–4–T

Applying Automatic Methods for the Construction of Multilingual WordNets

L. Benítez, G. Escudero, J. Farreres, G. Rigau

Índex

1 Introduction

In recent years the research in Natural Language Processing (NLP) field has proved the increasing importance of using extensive and complete Lexical Knowledge Bases (LKB's). So that problem had been approached by reusing, merging and tuning existing lexial material. In this way, the English Wordnet, developed at Princeton University by George Miller and its research group [Miller 90], has become a de-facto standard for the lexical semantic representation of the natural language for English.

WordNet success has encouraged several projects in order to build WordNets (WNs) for other languages, such as Catalan [Benítez et al. 98b], Spanish, Basque and Galician, for instance. Besides, multilingual links are established automatically through the connections with the English WordNet.

This report tries to summarize our approach to building miltilingual WNs, within the projects EWN [1], ITEM [2] and CREL [3]. (Centre de referència en Enginyeria Lingüística)

The main methodological approach followed up to date for the automatic construction of multilingual WNs (Spanish and Catalan) is described in the paper [Atserias et al. 97].

That paper includes the description of some class methods based on monosemic and polysemic criteria. Here we are going to describe a concrete algorithm to apply that class methods.

2 Previous conventions

Before describing how to apply the class methods, we are going to give some conventions for giving a compact description for the algorithm:

1. From the list of sets of *English words*, *non English words*, *synsets* and links between them that we have automatically collected from a bilingual dictionary [Benítez et al 98a],we build all pairs of *non English words* (NW) and *English words* (EW), and drop out duplicates. We denote that set as $Dict$.

[1] EuroWordNet LE4003
[2] ITEM TIC 96-1234-C03-03
[3] CREL CIRIT

A sample of *Dict*:

abacus àbac#m	abduction rapte#m
abandon abandó#m	ability aptitud#f
abandonment abandó#m	ability capacitat#f
abattoir escorxador#m	ability competència#f
abbess abadessa#f	ability habilitat#f
abbey abadia#f	ability talent#m
abbot abat#m	ability traça#f
abbreviation abreviatura#f	abode domicili#m
abdomen abdomen#m	abode habitacle#m

2. We consider the set of monosemous and polysemous English words relative to WN1.5. We denote them as $mono_{WN}$ and $poly_{WN}$ respectively.

A sample of $mono_{WN}$:

abalone	01288408
abamp	08344726
abampere	08344726
abandoned_infant	06098838
abandoned_person	05912614
abandoned_ship	02038160
abashment	04802547
abatis	02720878
abattis	02720878
abattoir	02038265
abbacy	05403491
abbe	05912735
abbe_condenser	02038397
abbess	05912800
abbot	05912913
abc	04340907

A sample of $poly_{WN}$:

abacus	02037873
abacus	02038006
abandon	03425896
abandon	04786649
abandonment	00027945
abandonment	00051049
abandonment	00116137
abasement	00152998
abasement	08725485
abatement	00203359
abatement	04730118
abbey	02038520
abbey	02038601
abbey	02038681
abbreviation	00201978
abbreviation	04586423

3. From *Dict* we consider the set of all pairs such which their non English word are monosemous relative to *Dict*. We denote it as $mono_{NE}$. In the same way we consider $mono_{EN}$, $poly_{NE}$, $poly_{EN}$.

A sample of $mono_{NE}$:

abacus	àbac#m
abbess	abadessa#f
abbey	abadia#f
abbot	abat#m
abbreviation	abreviatura#f
abdomen	abdomen#m
abruptness	brusquedat#f
absence	absència#f
absurdity	absurd#m
absurdity	absurditat#f
abundance	abundor#f
abuse	maltractament#m

A sample of $mono_{EN}$:

abacus	àbac#m
abandonment	abandó#m
abandon	abandó#m
abattoir	escorxador#m
abbess	abadessa#f
abbey	abadia#f
abbot	abat#m
abbreviation	abreviatura#f
abdomen	abdomen#m
abduction	rapte#m
abortion	avortament#m
absorption	absorció#f

A sample of $poly_{NE}$:

abandó#m	abandon
abandó#m	abandonment
abandó#m	forlornness
abandó#m	neglect
abandó#m	withdrawal
abast#m	compass
abast#m	extent
abast#m	range
abast#m	reach
abast#m	scope
abellot#m	bumblebee
abellot#m	drone

A sample of $poly_{EN}$:

ability	aptitud#f
ability	capacitat#f
ability	competència#f
ability	habilitat#f
ability	talent#m
ability	traça#f
abode	domicili#m
abode	habitacle#m
abruptness	brusquedat#f
abruptness	sequedat#f
abstract	extracte#m
abstract	resum#m

4. We denote the join operation with the symbol $\bowtie$.

5. With the symbol $\mathcal{N}(\mathcal{S})$ we denote the subset of *non English words* from any set $\mathcal{S}$.

A sample of $poly_{EN}$:

ability	aptitud#f
ability	capacitat#f
ability	competència#f
ability	habilitat#f
ability	talent#m
ability	traça#f
abode	domicili#m
abode	habitacle#m
abruptness	brusquedat#f
abruptness	sequedat#f
abstract	extracte#m
abstract	resum#m

A sample of $\mathcal{N}(poly_{EN})$:

aptitud#f
capacitat#f
competència#f
habilitat#f
talent#m
traça#f
domicili#m
habitacle#m
brusquedat#f
sequedat#f
extracte#m
resum#m

6. With the symbol $\mathcal{E}(\mathcal{S})$ we denote the subset of *English words* from any set $\mathcal{S}$.

A sample of $poly_{NE}$:

abandó#m	abandon
abandó#m	abandonment
abandó#m	forlornness
abandó#m	neglect
abandó#m	withdrawal
abast#m	compass
abast#m	extent
abast#m	range
abast#m	reach
abast#m	scope
abellot#m	bumblebee
abellot#m	drone

A sample of $\mathcal{E}(poly_{NE})$:

abandon
abandonment
forlornness
neglect
withdrawal
compass
extent
range
reach
scope
bumblebee
drone

3 General criteria

With the considerations from section 2 in mind, we describe how to build the disjunct sets of pairs of *English–non English* words (*EW–NW*):

Criterion 1:

A non English word *NW* has only one English translation; simmetrically, an English word *EW* has *NW* as its unique translation.

$$mono_{NE} \bowtie mono_{EN}$$

abacus	àbac#m
abbess	abadessa#f
abbey	abadia#f
abbot	abat#m
abbreviation	abreviatura#f
abdomen	abdomen#m
absolute_majority	majoria_absoluta#f

Criterion 2

A non English word *NW* has more than one translation; each English word *EW* has *NW* as its unique translation.

$$(\mathcal{N}(poly_{NE}) - \mathcal{N}(poly_{EN})) \bowtie poly_{NE}$$

abraçada#f	embrace
abraçada#f	hug
absorció#f	absorption
absorció#f	takeover
afluent#m	affluent
afluent#m	tributary

Criterion 3

Several *NWs* have the same translation; *EW* has several translations to *NW*.

$$(\mathcal{E}(poly_{EN}) - \mathcal{E}(poly_{NE})) \bowtie poly_{EN}$$

absurdity	absurd#m
absurdity	absurditat#f
acorn	aglà#mf
acorn	gla#f
acrobat	acròbata#m
acrobat	acròbata#mf

Criterion 4

Several NWs have the different translations; EW also has several translations

$$\mathcal{N}(poly_{NE} \bowtie poly_{EN}) - Dict \bigcup \mathcal{E}(poly_{NE} \bowtie poly_{EN}) - Dict$$

larva#f	larva
larva#f	grub
cuc#m	grub
cuc#m	worm
bestiola#f	worm
bestiola#f	insect
cuca#f	insect
cuca#f	bug
error#m	bug

Applying each of the four criteria described above, a distinct set of pairs formed by EW-NW is obtained. Next step lies in dividing each set in two different subsets, one containing the pairs with English monosemous words relative to WN1.5 and the other containing the polysemous ones. Finally we obtain 8 sets, each set fits one of the class methods (monosemic 1–4 and polysemic 1–4) from [Atserias et al. 97].

4 Results

Each method has been manually inspected in order to mesure its confidence score. Such tests have been manually performed on a representative random sample. The reliability obtained from the sample is assigned to the whole set. In table 1 are shown the results obtained for Spanish nouns, and in table 2 the ones for Catalan nouns.

CRITERION	#LINKS	#SYNSETS	#WORDS	%OK
mono1	3697	3583	3697	92
mono2	935	929	661	89
mono3	1863	1158	1863	89
mono4	2688	1328	2063	85
poly1	5121	4887	1992	80
poly2	1450	1426	449	75
poly3	11687	6611	3165	58
poly4	40298	9400	3754	61

Table 1: Overall results for Spanish nouns

CRITERION	#LINKS	#SYNSETS	#WORDS	%OK
mono1	1226	1212	1221	96
mono2	419	337	258	98
mono3	448	208	396	93
mono4	3012	1532	2178	94
poly1	2298	2244	864	90
poly2	568	519	158	78
poly3	1125	477	357	72
poly4	37714	9151	4266	55

Table 2: Overall results for Catalan nouns

5 Source code

```
gunzip -c wn_dict.noms.gz | gawk 'NF==3{print $1,$2}' | sort -u > dict

 # CRITERI 1

gawk 'BEGIN{prev="";line="";sum=0}$1==prev{line = line " " $2;next};
$1!=prev{if(line!="")printf "%s %s\n",prev,line; line=$2;prev=$1;}
END{printf "%s %s\n",prev,line;}' dict | gawk 'NF==2' | sort -u
| tr ' ' '|' | sort -u > monoEC

gawk '{print $2,$1}' dict | sort -u | gawk 'BEGIN{prev="";line="";sum=0}
  $1==prev{line = line " " $2;next};
  $1!=prev{if(line!="")printf "%s %s\n",prev,line;line=$2;prev=$1;}
  END{printf "%s %s\n",prev,line;}' | gawk 'NF==2' | sort -u |
  gawk '{print $2,$1}' | tr ' ' '|' | sort -u > monoCE

join monoCE monoEC | tr '|' ' ' | sort -u  > criteri.1

# PREVIS

gawk 'BEGIN{prev="";line="";sum=0}$1==prev{line = line " " $2;next};
  $1!=prev{if(line!="")printf "%s %s\n",prev,line; line=$2;prev=$1;}
  END{printf "%s %s\n",prev,line;}' dict
  | gawk 'NF>2'
  | gawk '{ for ( i = 2; i <= NF; ++i) printf "%s %s\n", $1, $i }'
  | sort -u > polyEC

gawk '{print $2,$1}' dict | sort -u
  |gawk 'BEGIN{prev="";line="";sum=0}$1==prev{line = line " " $2;next};
  $1!=prev{if(line!="")printf "%s %s\n",prev,line; line=$2;prev=$1;}
```

```
  END{printf "%s %s\n",prev,line;}' | gawk 'NF>2'
  | gawk '{ for ( i = 2; i <= NF; ++i) printf "%s %s\n", $1, $i }'
  | sort -u > polyCE

gawk '{print $2}' polyEC | sort -u > polyEC.cat
gawk '{print $1}' polyEC | sort -u > polyEC.ang
gawk '{print $1}' polyCE | sort -u > polyCE.cat
gawk '{print $2}' polyCE | sort -u > polyCE.ang

# CRITERI 2

join -v 1 polyCE.cat polyEC.cat | join - polyCE > criteri.2

# CRITERI 3

join -v 1 polyEC.ang polyCE.ang  | join - polyEC > criteri.3

# CRITERI 4

tr ' ' '|' < polyEC | sort -u > polyEC.sep
gawk '{print $2,$1}' polyCE | tr ' ' '|' | sort -u > polyCE.sep

join polyCE.sep polyEC.sep | sort -u | tr '|' ' '  > K

gawk '{print $1}' K | sort -u > K.ang
gawk '{print $2}' K | sort -u > K.cat

gawk '{print $2,$1}' dict | sort -u | join - K.cat | sort -u > criteri4a

join K.ang dict | sort -u > criteri4b

gawk '{print $2,$1}' criteri4a  | sort -u |  cat - criteri4b
  | sort -u > criteri.4

gunzip -c syns_word_n.gz | gawk '{print$2,$1}' | sort -u
  |gawk 'BEGIN{prev="";line="";sum=0}$1==prev{line = line " " $2;next};
  $1!=prev{if(line!="")printf "%s %s\n",prev,line; line=$2;prev=$1;}
  END{printf "%s %s\n",prev,line;}' | gawk 'NF==2' | sort -u > wn.mono

gunzip -c syns_word_n.gz | gawk '{print$2,$1}' | sort -u
  |gawk 'BEGIN{prev="";line="";sum=0}$1==prev{line = line " " $2;next};
  $1!=prev{if(line!="")printf "%s %s\n",prev,line; line=$2;prev=$1;}
  END{printf "%s %s\n",prev,line;}' | gawk 'NF>2'
  | gawk '{ for ( i = 2; i <= NF; ++i) printf "%s %s\n", $1, $i }'
```

```
  | sort -u  > wn.poly

join criteri.1 wn.mono > mono.criteri.1
join criteri.1 wn.poly > poly.criteri.1

gawk '{print $2,$1}' criteri.2 | sort -u | join - wn.mono > mono.criteri.2
gawk '{print $2,$1}' criteri.2 | sort -u | join - wn.poly > poly.criteri.2

join criteri.3 wn.mono > mono.criteri.3
join criteri.3 wn.poly > poly.criteri.3

join criteri.4 wn.mono > mono.criteri.4
join criteri.4 wn.poly > poly.criteri.4
```

References

[**Atserias et al. 97**] J. Atserias, S. Climent, J. Farreres, G. Rigau, H. Rodríguez. *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets* – Proceedings of Conference on Recent Advances on – NLP. RANLP 97. Tzigov Chark, Bulgaria, 1997.

[**Benítez et al. 98a**] L. Benítez, G. Escudero, J. Farreres, G. Rigau. *Converting the "Enciclopédia Catalana" bilingual MRD to an MTD.*– Technical Report LSI-98-3-T. – LSI Department. Universitat Politècnica de Catalunya, 1998.

[**Benítez et al. 98b**] L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, M. Taulé (Universitat Politècnica de Catalunya; Universitat de Barcelona) *Methods and tools for building the Catalan WordNet.* – In workshop on Language Resources for European Minority Languages at Conference on Language Resources and Evaluation (LREC).– Granada, Spain, 1998.

[**Miller 90**] G. Miller. *Five papers on WordNet.*–In special Issue of International Journal of Lexicography: actas 1990.