Some strategies for the improvement of a Spanish WordNet

Matías Herrera
Javier González
Luis Chiruzzo
Dina Wonsever

Facultad de Ingeniería
Universidad de la República
Uruguay

Global WordNet Conference 2016
Bucharest
Romania

Introduction

MCR (Multilingual Central Repository) contains WordNet versions for several languages spoken in Spain

- Spanish, Galician, Catalan, Basque

The synsets structure and relations are mostly the same as in English WordNet

The lemmas are translated into the different languages

Introduction

But its coverage is not as good as WordNet in English:

	Princeton WordNet	Spanish MCR
nouns	82k	39k
verbs	14k	11k
adjectives	18k	7k
adverbs	3.6k	1k

Coverage over Corin corpus

Coverage of Spanish MCR

	Lemmas in corpus	Lemmas in Spanish MCR
noun	4012	2780 (69.29%)
verb	1639	1235 (75.35%)
adjective	1647	840 (51.00%)
adverb	369	121 (32.79%)

Translation sources

Dictionaries:

- Wiktionary
- Apertium Eglish-Spanish dictionary

MT:

- Bing translator

Coverage over Corin corpus

Coverage of translation sources

	Translation is available	Translation available and present in English WordNet
noun	3529 (87.96%)	3108 (77.47%)
verb	1344 (82.00%)	1172 (71.16%)
adjective	1300 (78.93%)	1061 (64.73%)
adverb	309 (83.74%)	271 (73.44%)

Initial Selectors

Monosemy: If a lemma appears in only one English synset, use all its translations for the Spanish synset

Example:

- Synset eng-30-00048268-r has lemma “currently” in English
- Translations of currently: “hoy”, “ahora”, “actualmente”
- Add the three translations as lemmas for spa-30-00048268-r

Initial Selectors

Single translation: If a lemma has only one Spanish translation, use that translation in all synsets that contain the lemma

Example:

- Synsets eng-30-07328756-n, eng-30-10292316-n and eng-30-10480018-n have the lemma “producer”
- The only translation for “producer” according to the sources is “productor”
- Assign “productor” to synsets spa-30-07328756-n, spa30-10292316-n and spa-30-10480018-n

Initial Selectors

Factorization:

- Get all possible translations for all the lemmas of a synset
- Assign to the Spanish synset the translations that are shared by all the translation sources

Initial Selectors

Factorization:

Example:

- Synset eng-30-01309991-a has lemmas “artless” and “ingenuous”
- Translations for “artless”: “inocente”, “ingenuo”, “cándido”
- Translations for “ingenuous”: “inocente”, “ingenuo”
- Add “inocente” and “ingenuo” as lemmas for spa-30-01309991-a

Initial Selectors

Derived adverb:

- Find the adjective related to an adverb
- Find the Spanish translation of the adjective
- Use morphological rules to derive a Spanish adverb from the adjective
- Use a corpus to test if the derived adverb actually exists

Initial Selectors

Derived adverb:

Rules:

- Ends with “o” → Change the “o” to suffix “amente” (“lento” → “lentamente”)
- Ends with “r” or “n” → Add suffix “amente” (“encantador” → “encantadoramente”)
- Else → Add suffix “mente” (“legal” → “legalmente”)

Initial Selectors

Derived adverb:

Example:

- Synset eng-30-00033562-r (“mildly”) is linked to synset eng-30-01508719-a (“mild”)
- “mild” has translations “suave” and “leve”, so we derive adverbs “suavemente” and “levemente” (both exist in the corpus)
- “suavemente” and “levemente” are set as lemmas for spa-30-00033562-r

Evaluation - Initial Selectors

We chose 1000 synsets for each POS

We run the initial selectors on them

In total there are 6220 English lemmas

	Synsets that got a translation	Translation not present in MCR
noun	637	423 (54.8%)
verb	528	390 (50.4%)
adjective	599	429 (62.5%)
adverb	739	694 (83.8%)
total	2845	1936 (63.3%)

Evaluation - Initial Selectors

Precision of the initial selectors

POS	Monosemy	Single translation	Factorization	Derived adverb
noun	93.48%	100.00%	100.00%	
verb	93.48%	98.39%	100.00%	
adjective	96.08%	100.00%	96.00%	
adverb	97.14%	94.59%	92.00%	92.00%
total	95.04%	98.25%	97.00%	

Term Frequencies

Out of the 4000 synsets that were tested

2845 got a translation using the initial selectors

115 were not translated because no translation was found

1040 were not translated because they were ambiguous: more than one translation available

→ how do we disambiguate?

Term Frequencies

SA is hypernym of SB

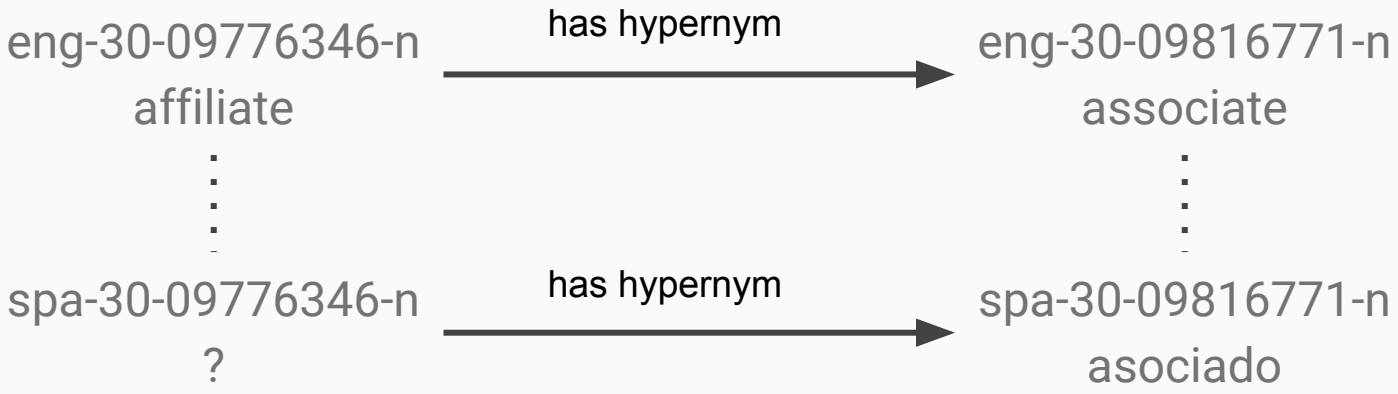
LA is an English lemma for SA, LB is an English lemma for SB

TB is a Spanish lemma for SB

We have two possible translations for LA: TA_1 and TA_2

Term Frequencies

Example:



Translation candidates: “filial”, “afiliado”

Term Frequencies

Θ(T) is the number of times the term T appears in a corpus

Θ(T_1,T_2) is the number of times T_1 and T_2 appear in the same sentence

For each candidate translation TA_i we calculate:

$$O_i = \frac{\Theta(TA_i, TB)}{\Theta(TA_i) + \Theta(TB)}$$

We choose the translation candidate TA_i which has the highest O_i

Term Frequencies

Example:

$O_{filial} = 0.0$

$O_{afiliado} = 8.18x10^{-5}$

So “afiliado” is taken as the translation of “affiliate” for synset spa-30-09776346-n

Term Frequencies

	Examples	Has relation	Has relation and translation
noun	230	230	179
verb	500	494	442
adjective	168	165	107
adverb	142	24	9
total	1040	913	737

Evaluation - Term Frequencies

	Examples with results	Not present in Spanish MCR
noun	127	79 (62.2%)
verb	282	206 (73.0%)
adjective	65	34 (52.3%)
adverb	3	1 (33.3%)
total	477	321 (67.3%)

Evaluation - Term Frequencies

	Precision
noun	68.0%
verb	68.0%
adjective	84.0%
adverb	100.0%
total	74.0%

Evaluation - Coverage over Corin corpus

	Total in corpus	Original coverage	Our coverage
noun	4012	2780 (69.29%)	2812 (70.09%)
verb	1639	1235 (75.35%)	1268 (77.36%)
adjective	1647	840 (51.00%)	895 (54.34%)
adverb	369	121 (32.79%)	232 (62.87%)

Conclusions

We tested a series of processes (selectors) that find new translations for WordNet synsets into Spanish

The initial selectors show promising results finding new translated lemmas, especially for adverbs

The term frequency selector also showed promising results, but its precision is lower

Future work

It was only applied to 1000 synsets for each category, we must apply it to the whole collection of synsets

The term frequency selector could be iterated because in each iteration it covers new synsets so there’s more available information

- we eventually run out of useful relations
- we need to improve precision first as each iteration will be less precise than the previous one

Thank you!

Mulţumesc!