Spanish WordNet 1.6: Porting the Spanish WordNet across Princeton versions

Jordi Atserias*, **Luís Villarejo***, **German Rigau**†

* TALP Research Center
Universitat Politécnica de Catalunya. Catalonia
{batalla, luisv}@talp.upc.es

†IXA Group, University of the Basque Country
Computer Languages and Systems
rigau@si.ehu.es

Abstract

This paper describes the new Spanish Wordnet aligned to Princeton WordNet1.6 and the analysis of the transformation from the previous version aligned to Princeton WordNet1.5. Although a mapping technology exists, to our knowledge it is the first time a whole local wordnet has been ported to a newer release of the Princeton WordNet.

1. Introduction

Using large-scale lexico-semantic knowledge bases (such as WordNet, Mikrokosmos, Cyc, etc.) has become a usual, often necessary, practice for most current Natural Language Processing systems. Building appropriate resources of this nature for open domain semantic processing is a hard and expensive task involving large research groups during long periods of development. For example, dozens of person-years are being invested world-wide into the development of wordnets for various languages.

Unfortunately, the outcomes of these projects are, usually, large and complex semantic structures, hardly compatible with resources released by other projects and efforts. Obviously, this fact has severely hampered language technology development. Thus, it is fundamental to have a robust, and highly accurate methodology to maintain the compatibility between lexical knowledge bases of different developers, languages and versions, past and new.

Although the technology to produce robust and accurate mappings between WordNet versions exists (Daudé et al., 2001), no complete methodology has been provided to produce completely new and upgraded versions of the local wordnets aligned to a particular WordNet.

2. The Spanish WN

The current version of the Spanish WordNet is the result product of ten years of combined effort of several research centers involved in different national and international projects. The first version of Spanish WordNet was build during the EuroWordNet project (Vossen, 1998). The Spanish WordNet construction followed the expand model. That is, following an automatic method and exploiting several Spanish-English bilingual dictionaries, WordNet synsets were translated into equivalent synsets in Spanish. In that way, an aligned version of WordNet 1.5 was built. This preliminary version was then corrected and augmented manually.

3. Porting Spanish WN to WN1.6

Now, the Spanish WordNet is being enhanced by the MEANING project. MEANING is a UE-funded project (IST-2001-34460) (Rigau et al., 2002) which has as one of its major goals the integration of several large-scale knowledge resources. MEANING has designed a Multilingual Central Repository (MCR) (Atserias et al., 2004) to act as a multilingual interface for integrating and distributing all the knowledge acquired in the project. The MCR will ensure the consistency and integrity of all the semantic knowledge produced by the project. The MCR follows the model proposed by the EuroWordNet project, whose architecture includes the Inter-Lingual-Index (*ILI*), a Domain ontology and a Top Concept ontology (Vossen, 1998).

The current version of the MCR integrates into a common framework:

- The ILI based in WN1.6, including the EWN Base Concepts, the EWN Top Concept ontology, MultiWordNet Domains (Magnini and Cavagli, 2000), Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001).
- Local WNs connected to the ILI, English 1.5, 1.6, 1.7, 1.7.1, 2.0, Basque, Catalan, Italian and Spanish WN.
- Large collections of semantic preferences, acquired both from SemCor and from BNC.
- Instances, including named entities.

To date, most of the knowledge uploaded into the MCR has been derived from data linked to WN1.6. However, the Spanish, Catalan and Basque WNs were aligned to WN1.5. To deal with the gaps between versions and to minimize side effects with other international initiatives (Balkanet, EuroTerm, eXtended WN) and WN developments around Global WordNet Association, we used a set of robust and accurate mappings between all involved English WNs to maintain the compatibility across wordnets.

Having all this WNs connected through the ILI, the knowledge acquired for each language can be consistently uploaded and integrated into the respective local WN, and then ported and distributed across the rest of WNs, balancing resources and technological advances across languages.

In particular, using these mappings we are able to port the entire local WNs to upgraded versions of the Princeton WN. There are multiple advantages in upgrading an existing local wordnet to a newer Princeton version: better structure, better coverage, connection with other resources.

This paper presents an exhaustive analysis of the porting process of the Spanish WN from WN1.5 to WN1.6. This analysis could help other WN developers to keep their local WNs up to date with respect to the latest Princeton wordnet.

4. Mapping Procedure

Uploading local wordnets based on WordNet1.5 to the MCR (aligned to WordNet 1.6) is a complex process, because between different wordnet versions, synsets can be splited (1:N), joined (N:1), added (0:1) or deleted (1:0) through mapping.

Regarding English versions, table 1 shows for each different casuistic of the mapping, the number of links between wn1.5 and wn1.6 synsets and the number of different synsets of each version involved. When applying the mapping to the Spanish wordNet the same casuistic applies (see table 2 for a summary), resulting in the losing of 346 synsets.

Even if we perform manual checking of these connections, for those remaining cases of splitting or joining synsets the information inside the synsets should be modified accordingly.

The whole process of the porting wordnets (synsets and relations) using ILI based on WordNet1.5 to the new ILI based on WordNet1.6 consist of:

- **Synsets** While *"local"* synsets (those created in the Spanish Wordnet) do not vary. The synsets coming from wn1.5 were transformed to wn1.6 as follows:

  1. For all splited synsets, all information of synset 1.5, including variants, is copied to each of the equivalent synsets in 1.6
  2. For all joined synsets, all information of synsets 1.5, including variants, is copied to the equivalent synset in 1.6
  3. Manual revision to validate the splitted and joined synsets.

- **Relations** As Spwn was build semi-authomatically from wn1.5, we consider that we should remove all the relation imported from wn1.5 and replace them with the relations coming from wn1.6. No relation coming from wn1.5 was passed through the mapping. Only those relation added with respect to wn1.5 were passed through. Through the mapping a relation can be joined, splited or multiplied. While, a pair of relations could be (equal or change the source or target synset). Even, we have similar cases for synsets but the impact in the Spanish wordnet is minimum. Thus, this paper will focus on the mapping of the synset.

	POS	**#links**	**#syn. wn1.5**	→	**#syn. wn1.6**
1:0	*noun*	-	530	→	-
	verb	-	160	→	-
	adj	-	243	→	-
	adv	-	33	→	-
	total	-	966	→	-
1:1	*noun*	65,740	65,740	→	65,740
	verb	10,841	10,841	→	10,841
	adj	17,824	17,824	→	17,824
	adv	2,854	2,854	→	2,854
	total	97,259	97,259	→	97,259
1:N	*noun*	69	34	→	69
	verb	42	21	→	42
	adj	171	83	→	171
	adv	30	15	→	30
	total	312	312	→	87
0:1	*noun*	-	-	→	4,994
	verb	-	-	→	964
	adj	-	-	→	2,440
	adv	-	-	→	448
	total	-	-	→	8,846
M:1	*noun*	683	683	→	338
	verb	212	212	→	106
	adj	1,374	1,374	→	665
	adv	168	168	→	81
	total	2,437	2,437	→	1,190
M:N	*noun*	8	4	→	4
	verb	4	2	→	2
	adj	2	2	→	1
	adv	0	0	→	0
	total	14	8	→	7

Table 1: Mapping wn1.5 → wn1.6 for Princeton WordNet

	POS	**#links**	**#syn. wn1.5**	→	**#syn. wn1.6**
1:1	*noun*	37,704	37,704	→	37,704
	verb	8,722	8,722	→	8,722
	adj	13,970	13,970	→	13,970
1:N	*noun*	57	28	→	57
	verb	28	14	→	28
	adj	167	81	→	167
N:1	*noun*	468	468	→	284
	verb	185	185	→	101
	adj	1311	1311	→	656
M:N	*noun*	6	3	→	4
	verb	2	1	→	2
	adj	2	2	→	1

Table 2: Mapping wn1.5 → wn1.6 figures for Spwn

5. Quality measures of the mapping

As a manual checking of the whole mapping or the resulting Spanish wordnet will be too time consuming, we will measure the quality of the mapping, measuring how much a synset (its contents and its relations with other synsets) has changed. The mapping divides the synsets in four categories according to the multiplicity of the mapping relation (1:1) (1:N) (M:1) (M:N). he mapping divides the synsets in four categories according to the multiplicity of the mapping relation (1:1) (1:N) (M:1) (M:N).

On one hand, we must perform a manual revision of all the cases where the mappings are not (1:1). For splitted synsets (*:N), because not all the information which is copied to each new resulting synset (i.e variants/glosses/relations) will be really shared by all of them. Similarly, for joined synsets (M:*), because the resulting content information will be repeated or not accurate.

Relaxation labeling algorithm tries to converge to the solution (mapping) that best hold a whole set of restrictions. Thus, the best selected mapping do not means that no changes has to be done in order to suit the new synset. Thus, even, in those cases where there is a 1:1 mapping we need to check the quality/consistency of the equivalences between English wordnet versions.

The quality of the mappings regarding its content can be measured by comparing the synonym set (exactly equal, contained, etc.) and glosses (empty, equal, included, overlap). Similarly, we measured also the changes in the relations of the synset through the mapping by measuring the changes in the WN relations. We choose a very simple way of combining the different measures by just adding their values. First, the quality measure for each mapping between a wn1.5 synsets an wn1.6 synsets is calculated. Then, the quality of the resulting wn1.6 synset (confidence score) is defined as the minimum of the quality of all its mappings.
The next section describes the set of measures used.

5.1. Variant Based Measures (QVariant)

This measure is based in the overlapping between the set of variants of the source and target synset.

Comparing the two set of variants we can find that:

- **EQ** Both set of variants are equal.
- **EXTENDED** Wn1.6 variants includes Wn1.5 variants.

synset	variants
00003128r	just#1 merely#1 only#1 simply#1
00003737r	but#1 just#1 merely#1 only#1 simply#1

- **REDUCED** Wn1.5 variants include Wn1.6 variants

synset	variants
00003345v	hiccough#1 hiccup#1 make_a_hiccup#1
00002841v	hiccough#1 hiccup#1

- **MIX** Consider the number of common variants

synset	variants
00022594r	almost#2 close_to#1
00006065r	about#1 approximately#1 around#5 close_to#1 just_about#2 more_or_less#1 or_so#1 roughly#1 some#1

We calculate the score as: twice the number of common variants divided by the number of variants. Note that due to the mapping construction there is at least a common variant.

5.2. Gloss Based Measures (QGloss)

This measure is based in the overlapping between the glosses of the source and target synset. Before comparing glosses, the gloss examples are removed. Obviulsy these measures can be considerably improved by lemmatising or parsing the glosses.

- **EQ** Both glosses are equal
- **NEAREQ** Both glosses are equal removing text inside parenthesis
- **EXTENDED** Wn1.5 gloss is a part of Wn1.6 gloss
- **REDUCED** Wn1.6 gloss is a part of Wn1.5 gloss
- **MIX** Consider the number of common words

synset	gloss
00005659a	being the most complete of its class
00005386a	being the most comprehensive of its class

- **NULL** There is no gloss in wn1.5.

We calculate the score as: twice the number of common words divided by the number of words. Zero if the method is not applicable and -1 if there are no common words.

5.3. Semantic File Measures (QSemf)

This measure is based in the overlapping between the semantic file of the synset in wn1.5 and the mapped wn1.6 synset. This measure scores 1 if equal and -1 otherwise (there are 431 differences).

5.4. Relationship Based Measures (Qrel)

While, to measure the change on the contents of the synsets we can only relay on the English wordNet information, regarding relations we can use the whole set of relation of the Spanish Wordnet (See table 6. for a summary of them).

Once all the wordnet1.5 synsets are mapped to wn1.6, the relations can be mapped accordingly. This measure is based in the overlapping between the set of relations of one synset in SpWn1.5 and their equivalent/s in SpWn1.6.

- **EQ**: All the Spwn1.5 relations have a corresponding Spwn1.6 relation.
- **CHANGED**: When some SpWn1.5 relations do not have a corresponding Spwn1.6 relation. Then, the quality is calulated as the relation kept in wn1.6 divided by the number of relations from SpWn1.5.
- **NONE**: None of the SpWn1.5 relations is kept.

Relations	**Number**
be_in_state	1,302
causes	240
has_derived	8,504
has_holo_madeof	708
has_holo_member	11,847
has_holo_part	6,878
has_subevent	427
has_xpos_hyponym	319
near_antonym	7,444
near_synonym	10,965
role	106
role_agent	516
role_instrument	291
role_location	83
see_also_wn15	3,280
xpos_fuzzynym	37
xpos_near_synonym	319
Total	53,272

Table 3: Summary of Spwn1.6 Relations

6. Results

We should point out that almost the half of the synsets (42,161) has exactly the same variants and gloss. Table 6. shows the quality per POS of the 1:1 mapping (which is equivalent to the quality of the synsets having a 1:1 mapping). A global quality measure of 0.88 means that the impact in the Spanish WordNet will be minimum. Even, verb glosses seems to be more conflictive than for the rest of POS and the relation measure is quite bad in adjectives (maybe because there are few of them).

POS	**QVar**	**QGloss**	**QSem**	**QRel**	**Quality**
noun	0.85	0.92	0.99	0.76	0.88
verb	0.91	0.77	0.99	0.92	0.90
adj	0.94	0.81	0.98	0.66	0.84
total	0.88	0.87	0.99	0.76	**0.88**

Table 4: Quality measure for 1:1 en15 Spanish synsets

Table 5 shows the figures of Spwn1.5, the number of synsets which comes from Princeton WN1.5, the number of *"local"* synsets, the resulting synsets aligned to Princeton WN1.6 after the mapping and the final figures for Spwn1.6. As we can observe there is no much changed in coverage.

As a result of upgrading the Spanish WordNet to WordNet 1.6 (UPLOAD), different information coming from resources aligned to WN1.6 (SUMO, IRST's domains, large collections of Selectional Preferences, Instance Information) are now available for the Spanish wordNet (see table 6) in the MEANING PORT0 (Atserias et al., 2004).

pos	**Spwn1.5**	**syn1.5**	**local**	**syn1.6**	**Spwn1.6**
noun	43,652	38,308	5,344	38,023	43,367
verb	9,258	9,045	213	8,830	9,043
adj	15,859	15,585	274	14,667	14,941
total	68,769	62,938	5,831	61,520	67,351

Table 5: figures for Spanish WordNet

Relations	**PORT0**
role_agent-semcor	+52,394
role_agent-bnc	+67,109
role_patient-semcor	+80,378
role_patient-bnc	+79,443
Role	+279,324
Instances	+1,599
Domains Synsets	+48,053

Table 6: Spwn1.6 gained relations

7. Conclusions

We have described the new Spanish Wordnet aligned to Princeton WN1.6. In the MEANING framework, the Spanish WordNet is being connected to several knowledge resources (SUMO, MWND) and enhanced with large collections of semantic preferences obtained from English. We also have carried out an exhaustive analysis of the porting process of the Spanish WN from WN1.5 to WN1.6. This analysis could help other WN developers to keep their local WNs up to date with latest Princeton wordnet versions.

Using a set of metrics we are not only able to measure the impact of a version transformation, but also can detect conflictive cases. Associating confidence scores to the information present in the Spanish Wordnet is crucial for MEANING. These measures are the base to perform a semi-automatic acquisition and integration of knowledge as well as a manual improvement of the whole wordnet. This process is still undergoing and actually covers about 50,000 wordnet senses.

8. References

Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen, 2004. The meaning multilingual central repository. In *Second International WordNet Conference-GWC 2004*. Brno, Czech Republic. ISBN 80-210-3302-9.

Daudé, J., L. Padró, and G. Rigau, 2001. A complete wn1.5 to wn1.6 mapping. In *Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*. Pittsburg, PA, USA.

Magnini, B. and G. Cavagli, 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*. Athens. Greece.

Niles, I. and A. Pease, 2001. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Chris Welty and Barry Smith, eds.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll, 2002. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of COLLING Workshop*. Taipei, Taiwan.

Vossen, P. (ed.), 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .