Selecting the Correct English Synset for a Spanish Sense

Javier Farreres, Horacio Rodríguez

Universitat Politècnica de Catalunya
Barcelona
farreres,horacio@lsi.upc.es

Abstract

This work tries to enrich the Spanish Wordnet using a Spanish taxonomy as a knowledge source. The Spanish taxonomy is composed by Spanish senses, while Spanish Wordnet is composed by synsets, mostly linked to English WordNet. A set of weighted associations between Spanish words and Wordnet synsets is used for inferring associations between both taxonomies.

1. Introduction

In previous work, a way to extract a large volume of weighted associations between Spanish words and WordNet (WN) synsets was obtained (Farreres et al., 2002). Its results derive from the entries of a Spanish-English bilingual dictionary and are thus limited by the coverage of the dictionary. In order to extend the associations to words not covered by the bilingual dictionary a taxonomy alignment was considered. This work is centered in the nominal part[1].

Main problems in ontology merging arise when 1) the units of the ontologies to be merged are of different granularity, and 2) the mapping between those units cannot be stated directly but through another intermediate level of representation. This is the case of lexico-conceptual taxonomy merging where the mapping between conceptual units is performed through intermediate lexical units. An even more complicated case occurs when the mapping between conceptual units is performed through two levels of intermediate lexical units. This latter case reflects the approach that was taken to create the Spanish Wordnet (SpWN) in the first stages of development within the EuroWordNet (EWN) project (Atserias et al., 1997). See figure 1.

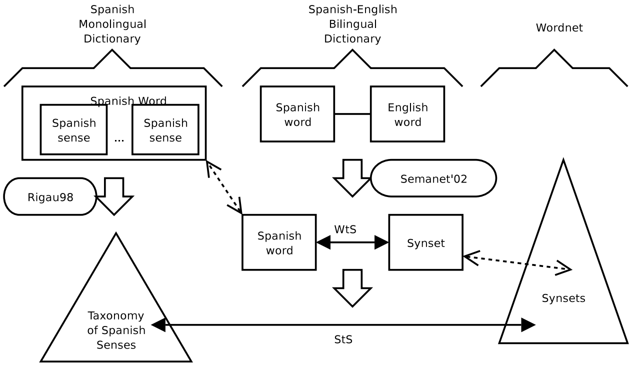


Figure 1: Alignment framework

Our aim in this paper is to use a Spanish sense-level taxonomy (SpTax), automatically extracted from a monolingual dictionary following (Rigau, 1998) for enriching the SpWN. In order to achieve it we use the English WN1.5, and the partially filled SpWN. There is room for enriching SpWN simply attaching Spanish Words to the yet uncovered English synsets. See table 1 columns 1, 2 for a comparison of the different volumes of the taxonomies involved.

	WN1.5	SpTax	WtS	StS
Words	87642	62433	12073	9509
Variants	107424			
Senses		111512		35566
Synsets	60557		18650	17443
Associations			65304	326368

Table 1: Taxonomy volume comparison

Mapping is allowed by means of Spanish words, that correspond to Spanish synsets linked to their corresponding WN synsets, and at the same time to Spanish senses in SpTax. The problem is that the correspondence between conceptual units (synsets in WN, senses in SpTax) results in a many-to-many case.

2. Terminology

We first introduce some concepts that will be used along the text. A *Spanish word* is a word covered by a monolingual dictionary. A *Spanish sense* is a sense of a word as defined by the monolingual dictionary, thus depending of the source. Two kinds of *associations* are considered. A *WtS* is a weighted association between a Spanish word and a WN synset; the *weight* of each association is the probability of correct assignment under the logistic model obtained in (Farreres et al., 2002), named in this paper as the *logistic probability* of the link. An *StS* is an association between a Spanish sense and a WN synset. See table 1 columns WtS, StS for a comparison of the volumes involved. The *branch* starting at some sense is the sequence of ancestors of that sense, including the sense. A *gap* in a Spanish branch is a sub-branch without associations (often reduced to a single node) separating two sub-branches with associations.

The PRB Given a *StS a*, *PRB(a) (pair of related branches)* is formed by the Spanish branch developed up to 5th ancestor, following the results in (Farreres et al., 2003), the WN branch developed up to the topmost level, and all the associations relating both branches. The *level* of an association in the *PRB* is the level of the Spanish ancestor from which the association starts. A *PRB* is said to be *connected* when some Spanish ancestor is associated with the WN branch. Connected *PRB*s have a *level*, which is the

[1] This research has been partially funded by Aliado (TIC2002-04447-c02-01).

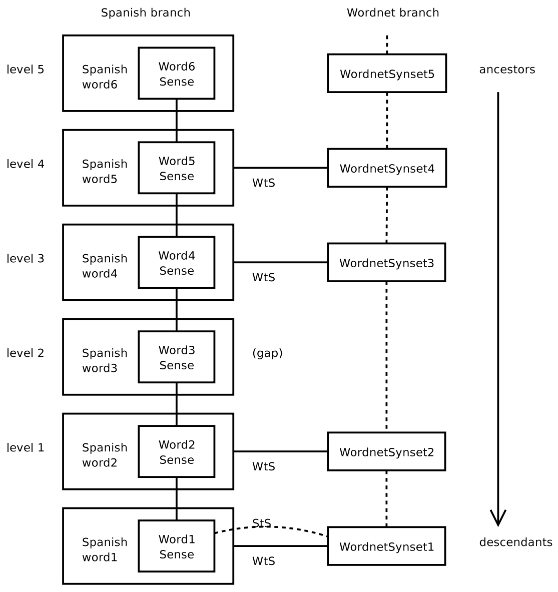


Figure 2: Diagram of a *PRB*

level of the first Spanish ancestor with an association in the *PRB*, the ancestor which is closest to the link originating the *PRB*

PRB is the concept managed in this work that allows studying the relationship between SpTax and WN.

3. Problem Definition

This work addresses the transformation of a set of *WtS* into a set of *StS*.

Given a *WtS*, the information the association provides is that the Spanish word is, probably, a translation of some variant occurring in the synset. But a word has not a unique meaning; meanings are in the senses of the words. Although knowing that a Spanish word may correspond to a WN synset is an useful information, the aim should be to assign the adequate sense of the Spanish word. As a side effect, this would also help detecting several wrong *WtS*, when for them no adequate Spanish sense were found.

Due to the complexity of the problem, a cautious approach has been preferred. A step by step analysis has been performed, starting from the analysis of simple cases, in order to first understand the kind of problems occurring in this type of taxonomy merging. A further step will consist on applying this knowledge to more complex cases.

4. Analysis of Monosemic Spanish Words

Following this incremental approach, from simple to complex, in a previous work (Farreres et al., 2003) we proposed a comparison between two taxonomies as a way to transform a set of *WtS* into a set of *StS*. The study was centered in the most simple case, monosemic Spanish words with only one *WtS*, converting it to an *StS* assigned to the single sense straightforwardly. 1263 *StS* where induced from *WtS* this way with a percentage of 96.7% of correctness. After building the whole branches of ancestors of both the Spanish sense and the synset of the association, it was first observed from the data that SpTax chains should be limited to five ancestors, as only few cases had their first association after this limit, and the correctness of an association this far was dubious. When the confidence scores of the associations calculated using the logistic regression model obtained in (Farreres et al., 2002) were studied, it was observed that the mean probability of the associations was related to the cardinality of the relation between the branches, the presence of gaps in the Spanish branch, and the number of Spanish ancestors with an association.

Group	Quantity	Percentage
No PRB connected	39	26%
One PRB connected	49	33%
Both PRB connected	62(*)	41%
Total	150	

(*)Two of them are tops in SpTax and are thus not taken into account in this work.

Table 2: Pairs of *PRB* for the same Spanish word

4.1. Monosemic Spanish Words with Two Associations

After analyzing the behavior of monosemic Spanish words with one *WtS*, the next step forward is to study what happens when a monosemic Spanish word has two or more *WtS*. In this case possibilities increase, as it can be the case that some *WtS* are incorrect and some are correct; the problem stands in the separation of the correct and incorrect *WtS* and the ulterior transformation into *StS*. In a first step the case where only two *WtS* are present is studied. At the end results are projected to the set of monosemic Spanish words with a larger number of associations.

Having the evidence that some factors are related to the probability of the base *StS* of *PRB*s, a study has been carried out on 150 monosemic Spanish words with two *WtS*, taking into account the sense and the two links. The factors studied for each link that seem to be relevant are: the existence of an ancestor with an association together with its level, the cardinality of the relation between the Spanish and the English branches, the presence of gaps in the Spanish branch, the number of Spanish senses with an association, the manual evaluation of the association, and the probability of the link.

The first distinction has been done in terms of whether the *PRB*s are connected. Table 2 summarizes the set volumes. When a *PRB* is connected, there is added evidence as to the correctness of the base *StS*.

4.1.1. No PRB Connected

When no *PRB* is connected, it makes no sense calculating the factors, as no upper association can be found in the *PRB*, and no *level* nor number of associations can be extracted as new evidences to be studied. Upon a detailed inspection, there were 4 cases where both *StS* were correct, 7 cases where both *StS* were incorrect, and 29 cases where one *StS* was correct. The only factor that can be contrasted is the logistic probability of the association, but no relation was detected between the logistic probability and the fact of an association of being correct.

In a more detailed study, 7 correct cases could be detected where the first ancestor of both branches shared ho-

Equivalent senses

	Estimated		
Real	yes	no	
yes	41 (42%)	13 (13%)	54 (55%)
no	8 (8%)	36 (37%)	44 (45%)
	49 (50%)	49 (50%)	98

Table 3: One *PRB* connected

Equivalent senses

	Estimated		
Real	yes	no	
yes	12 (46%)	2 (8%)	14 (54%)
no	1 (4%)	11 (42%)	12 (46%)
	13 (50%)	13 (50%)	26

Table 4: Both *PRB* connected and one *PRB* with greater *n* or lower *level*

mographs. As the current work only studies existing associations, this will not be taken into account right now. But in a future line of enrichment of SpWN, the generation of associations by means of homographs will be studied, with the caution of the danger of false friends.

4.1.2. One PRB Connected

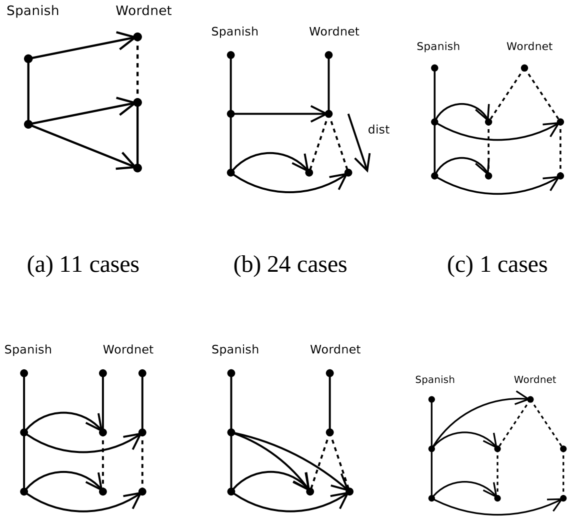
When only one of the *PRB* is connected, in most of the cases the connected link has proven to be correct, while the disconnected *PRB* is in mostly incorrect. The 49 cases generate 98 *PRB*: the 49 connected *PRB* have been accepted, the other 49 *PRB* have been rejected. Table 3 shows that the detection has a recall of 76% and a precision of 84%. The number of correct solutions discarded is quite high (13) but no evidence of any factor was detected that could help recovering any of them.

4.1.3. Both PRB Connected

When both *PRB* are connected, there is the possibility that both are correct (16 cases), that just one of them is correct (43 cases), or that both are incorrect (2 cases). The factors (*level* of upper closest association, *cardinality* of the relation between the branches, presence of *gaps*, number *n* of associations, *evaluation*, logistic *probability*) have been studied for each of the cases.

It was observed that the behavior of factors *level* and *n* is quite different when one or both *PRB* are correct. Upon studying them, it is observed that for the case that one *PRB* is incorrect, it is almost always the case that the correct *PRB* has higher *level* or higher *n* than the incorrect *PRB*. Applying this result, table 4 shows the results of separating correct and incorrect cases when the above condition takes place. The detection of correct and incorrect cases when one *PRB* has greater *n* or *level* than the other is quite precise,with a recall of 85.7% and a precision of 92%.

When both *PRB* have the same *level* and *n*, the structure of the association with the WN branch of the *PRB* has been studied. Six structures have been distinguished. Figure 3 shows the structures together with the number of cases of each of the structures; case b) is the most frequent totaling a 42%, and the second most frequent is case a) with a 19%.



(a) 11 cases (b) 24 cases (c) 1 cases

(d) 3 cases (e) 6 cases (f) 2 cases

Figure 3: Structures of *PRB*

Equivalent senses

	Estimated		
Real	yes	no	
yes	5 (42%)	1 (8%)	6 (50%)
no	1 (8%)	5 (42%)	6 (50%)
	6 (50%)	6 (50%)	12

Table 5: Algorithm results for structure b

The two most frequent structures covering a 61% of the cases have been studied more deeply. The rest has been left out of further study, and will be accepted as correct globally, mainly due to the lack of coverage.

In case a) the decision is between a synset and its immediate ancestor. In 3 cases both are correct, in 5 cases the lower one is the correct one, and in 3 cases the upper one is the correct one. No way to distinguish them automatically has been found, and evaluation will be left out for the manual validation process, accepting both as correct, giving 22 *StS* with a 64% precision.

In case b) a pattern has been detected that tends to distinguish the correct links from the incorrect ones in a small number of cases. A *distance* is defined counting the number of nodes between each of the base links generating the *PRB* and the lowest upper association common to both *PRB*, see figure 3 b). Given two *PRB* A and B, when |d(A)-d(B)|>2, the *PRB* with lower distance is usually correct and the other one incorrect. Table 5 shows the results, giving a recall of 83.3% and a precision of 83.3%.

Table 6 shows the results of the successful part of the study for monosemic words with two associations, with a recall of 78% and a precision of 85%. If the rest of unresolved *PRB* were to be accepted as correct globally, the results would return a 75% recall and 75% precision, being the main source of incorrect values accepted as correct the number of unresolved *PRB* where both are connected.

Equivalent senses

	Estimated		
Real	yes	no	
yes	58 (43%)	16 (12%)	74 (54%)
no	10 (7%)	52 (38%)	62 (46%)
	68 (50%)	68 (50%)	136

Table 6: Global results

Equivalent senses

	Estimated		
Real	yes	no	
yes	85 (15%)	40 (7%)	125 (22%)
no	179 (31%)	270 (47%)	449 (78%)
	264 (46%)	310 (54%)	574

Table 7: General case considering *PRB* connectness

This subgroup will need to be studied more deeply in future work.

4.2. Generalization to the Case of Many Associations

In the set of monosemic Spanish words there are 554 with 2 *WtS* and 865 with more *StS*. A sample of 48 monosemic Spanish words with more than one *WtS* has been extracted. The 48 Spanish words have 604 *WtS*, with a mean of 13 *WtS* per Spanish word, with the minimum being 4 *WtS*, and the maximum being 31 *WtS*. One of the words is in fact a top in SpTax, and thus its 30 associations are discarded. The resulting 574 *WtS*have been transformed into *StS* and their corresponding *PRB* have been constructed. The manual validation gives 125 correct, a 22%, and 449 incorrect. A large quantity of the incorrect *PRB* is the result of the polysemy of the English translations, originating erroneous associations. Another big set of incorrect *PRB* are correct *WtS* that deem incorrect when compared with the Spanish senses. In this case, the usual problem is that the bilingual Spanish-English dictionary gives a translation for a sense that is not represented in the monolingual Spanish dictionary.

Table 7 shows the results of the study. The detection of incorrect *PRB* has 87% precision and 60% recall. But the detection of correct *PRB* has a 32% precision and 68% recall. It seems from these results that the detection of incorrect cases works much better than the detection of correct results. This is identificative of semidecidible problems.

When adding the comparison of number of ancestors with association together with the comparison of distances of case b, the precision increases to 37% but the recall decreases to 47% (see table 8). In this case, whenever some *PRB* are related via structure a), they have been taken as one node in order to calculate distances; all have been assigned the same distance when compared with other *PRB* via structure b) in figure 3 .

5. Global Results and future work

A set of factors has been identified that help separating correct and incorrect *StS* starting from the same Spanish sense by evaluation and comparison of the generated *PRB*. Disconnected *PRB* are taken as incorrect. Connected *PRB* are compared in terms of the number of Spanish ancestors with association, the level of the first ancestor with an association, and the structure of the relation with WN. Centering the study on the set of Spanish words with 2 *WtS*, the successful measures return 81% recall and 85% precision. If the unsolved cases are included, the results return 75% recall and precision.

Equivalent senses

	Estimated		
Real	yes	no	
yes	59 (10%)	66 (11%)	125 (22%)
no	101 (17%)	348 (61%)	449 (78%)
	160 (29%)	414 (72%)	574

Table 8: General case applying comparisons

	Recall	Precision
Mono Spanish words with 1WtS	100%	96.7%
Mono Spanish words with 2WtS	81%	85%
Mono Spanish words more WtS	47%	37%
Mono Spanish words	67%	58%

Table 9: Global results for monosemic Spanish words

Translating the results to the case of monosemic Spanish words with more than 2 *WtS*, the separation has a precision of 37% with a recall of 47%. These figures are low, but it should be taken into account that the set of words under study is a quite complex one: monosemic Spanish words with more than 2 *WtS*. It is a suspicious behavior for a monosemic word, and logically it results in low figures. Table 9 summarizes the results for all monosemic Spanish words.

Table 7 shows that unconnected *PRB* are a good indicator of incorrect *StS*. In this study, *PRB* have been separated between connected and disconnected. But many of the connected *PRB* are supported by unconnected upper *PRB*. If a chain of deletions were executed along the branches, many of the incorrect *PRB* would disappear.

6. References

J. Atserias, S. Climent, J. Farreres, G. Rigau, and H. Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Tzigov Chark, Bulgaria.

J. Farreres, K. Gibert, and H. Rodríguez. 2002. Semiautomatic creation of taxonomies. In G. Ngai *et al.*, editor, *Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks'*, Taipei, August.

J. Farreres, K. Gibert, and H. Rodríguez. 2003. Towards binding spanish senses to wordnet senses through taxonomy alignment. In Sojka et al., editor, *Proceedings of the Second International Wordnet Conference (GWC 2004)*, pages 259–264, Brno. Masaryk University.

G. Rigau. 1998. *Automatic Acquisition of Lexical Knowledge from MRDs*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.