Towards Binding Spanish Senses to Wordnet Senses through Taxonomy Alignment

Javier Farreres[1], Karina Gibert[2], and Horacio Rodríguez[1]

[1] Computer Languages and Systems Department, Universitat Politècnica de Catalunya, Campus Nord, 08028 Barcelona, Spain, Email: farreres@lsi.upc.es
[2] Statistics and Operations Research Department, Universitat Politècnica de Catalunya

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 Wordnet is composed by synsets (English senses). A set of weighted associations between Spanish words and Wordnet synsets is used for inferring associations between both taxonomies.[3]

1 Introduction and Previous Work

This work continues a line of research directed to build Wordnets for languages in an automated way. Trying to delay human intervention as much as possible, a taxonomy alignment is performed. Using a set of associations previously obtained between Spanish words and Wordnet synsets, together with a logistic model that weights those associations, and a Spanish taxonomy of senses extracted with automatic processes, inference of associations from Spanish senses to Wordnet synsets is studied. This work uses results obtained in some previous works, introduced below.

In [Atse98] the interaction between different methods that link Spanish words with WordNet synsets was studied. Intersections between pairs of methods were proposed for maximizing the number of links together with the global accuracy.

In [Farr02] a methodology considering maximal intersections among all the methods in [Atse98] was proposed. A logistic model[4] was obtained for estimating the probability of correctness of a given link.

In [Rigau98] a method is offered for automatically generating a taxonomy of Spanish senses by means of a Word Sense Disambiguation process on the genus of the dictionary definitions. Even though the genus is detected with adequate precision, the sense discrimination has a much higher degree of error. This causes that, when building a complete branch of Spanish senses from the taxonomy, at some point some error will deem a chain of incorrect ancestors.

In [Farr98] some ideas were proposed related to taxonomy alignment. Studying simple geometrical configurations that tie the Spanish taxonomy of [Rigau98] with Wordnet, ways to increase probability of selected associations were proposed, as well as possibilities for inferring new associations.

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

[4] The logistic regression approximates the probability associated with a vector of booleans.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 259–264.

2 Basic Concepts

The terms defined below are used along this paper.

A *Spanish word* is a word covered by a Spanish monolingual dictionary. A *Spanish sense* is a sense of a Spanish word as defined by the Spanish monolingual dictionary, thus it is source dependent. Two kinds of associations are considered. A *WtS* is an association of a Spanish word to a Wordnet synset, with a probability of correctness, named in this paper as the *logistic probability*, calculated with the logistic model obtained in [Farr02]. An *StS* is an association of a Spanish sense to a Wordnet synset. Whenever a Spanish sense has no *StS*, it may always inherit the *WtS* of the word it belongs to. The *branch* starting at some sense is the sequence of ancestors of that sense up to the top, including the sense. A *gap* in a Spanish branch is a Spanish sub-branch that has no association and that separates Spanish sub-branches with associations.

The PRB Given an *StS* c, *PRB(c)* (*pair of related branches*) is defined as the pair of branches developed upward, on the one hand, from the Spanish sense till sixth level (as justified in section 5) and, on the other hand, from the corresponding Wordnet synset up to the top, together with all the associations connecting both branches. See figure 1 for a graphical example.

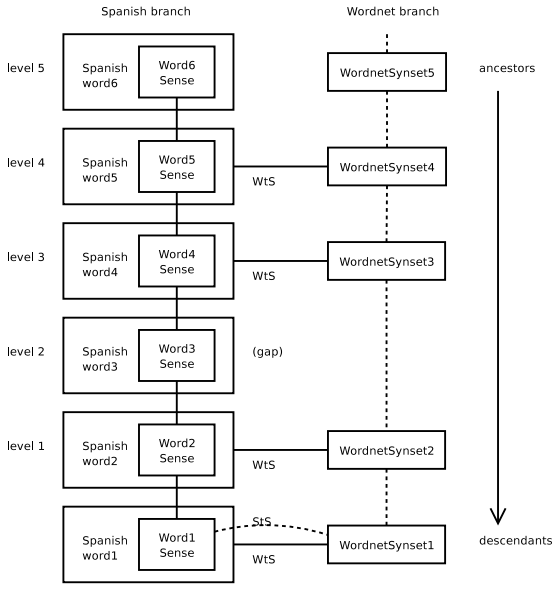


Fig. 1. The PRB

PRB is the concept managed in this work to allow study of the relationship between the Spanish taxonomy and Wordnet.

3 Towards the Induction of Upper Connections

After obtaining 66.000 associations in [Farr02], the only work left seemed to be a manual validation. But upon observing the data, many obviously wrong results were detected, frequently related to *WtS* of Spanish words without any Spanish sense supporting this association. If the Spanish senses could be contrasted with those *WtS*, the obvious errors could be deleted automatically. The natural resource to be applied, and the one that was at our reach, was a Spanish taxonomy of senses, even if generated automatically.

When the alignment of two taxonomies was considered, other useful applications arose as transforming *WtS* into *StS*, inferring *StS* without previous *WtS*, detecting erroneous *WtS* and knowing which Spanish senses remain uncovered.

4 Induction of Basic Connections

As a first stage of this research, monosemic Spanish words with only one *WtS* were considered. For this specific case, *StS* can be directly obtained from *WtS* to produce a starting kernel. Using the Spanish taxonomy as knowledge source, 1263 such monosemic Spanish words were detected, giving 1263 *StS* between 1263 Spanish senses and 1195 Wordnet synsets, that is, 1.06 Spanish senses per Wordnet synset.

From those, 685 *StS* were randomly chosen and manually evaluated, obtaining 559 correct evaluations and 19 incorrect evaluations giving a global accuracy of 96.7%.

5 The PRB Concept

Knowing that the Spanish taxonomy is not error free in the detection of the parent sense, the behavior of the Spanish ancestors of the senses taking part in the *StS* was studied for determining the distribution of errors.

For each *StS* the complete branch of the Spanish sense was built using [Rigau98], the complete branch of the synset was retrieved from Wordnet, and all the associations connecting both branches were identified.

Those parallel branches were classified on the basis of the level of the first association (a *WtS* or a previously identified *StS*) above the base *StS*, and the results are shown in table 1. The case where no ancestor has an association is the one that accumulates the highest number of errors, 14. In few cases the level is above five.

Upon these results, it was decided that further experiments would be carried out with Spanish branches of up to 5 ancestors, while the Wordnet branch would have no limit. The set of parallel branches within these parameters was named *PRB*, and the 685 *StS* were used to generate the corresponding *PRB*s.

6 PRB with Association on the First Level

298 of the *PRB*s generated have and immediate Spanish ancestor with an association above the base *StS*. For those *PRB*s, three parameters have been studied: the cardinality of the relationship between the branches in the *PRB* as defined by the set of associations, the number

Table 1. Level of first ancestor with an association

Level	Count	OK	KO	%
none	159	145	14	91
1	298	298		100
2	70	67	3	95
3	29	28	1	96
4	11	10	1	90
5	6	6		100
7	3	3		100
11	1	1		100
27	1	1		100

of senses with an association in the Spanish branch of the *PRB* and the existence of gaps in the Spanish branch.

The set of associations of each *PRB* defines a relationship between the two branches. The cases below were detected. Table 2:left shows the number of *PRB*s with structure in each of the cases.

***PRB* with crossings:** two Spanish senses have associations that cross each other (see cases *d)*, *e)* in figure 2). There are only 11 cases, probably due to errors in the sense disambiguation process or to differences of lexicalization between the two languages. It was not studied further.

1:1: only one association links any Spanish sense, and only one association links any synset.

1:N: only one association links any Spanish sense, but several associations may link any synset (see cases *a)* to *c)* in figure 2).

N:1: several associations may link any Spanish sense, but only one association links any synset (see cases *f)* to *i)* in figure 2).

M:N: several associations may link any Spanish sense, and several associations may link any synset.

For every *PRB*(c) the logistic probability of c was obtained and table 2:left shows the average probability per group. Groups 1:1 and 1:N have a similar mean probability, higher than the other two groups. It seems, then, that the structure of *PRB*(c) may provide some useful information about the correctness of *StS* c.

Continuing with the study of the probability of c, the number of Spanish senses with an association in the *PRB* was proved to be related to the value. However, a third factor, the existence of gaps or not, demonstrated to affect the relation. In table 2:right the mean probability of *StS* c depending on the number of associations of *PRB*(c) and the existence of gaps or not is displayed. It can be seen how the mean probability increases with the number of associations, and also it is greater if no gaps exist.

The behavior of the mean probability taking into account the three factors together is displayed in table 2:center, which shows that the correctness of c in general increases with the number of Spanish senses with an association in the *PRB* without gaps and also if it presents an 1:1 or 1:N structure. However the behavior of *PRB*s with gaps is less clear in this context.

Fig. 2. PRB configurations for classes 1:N, N:1, with crossings
Solid lines mean direct relations, dotted lines mean indirect relations, arrows mean associations

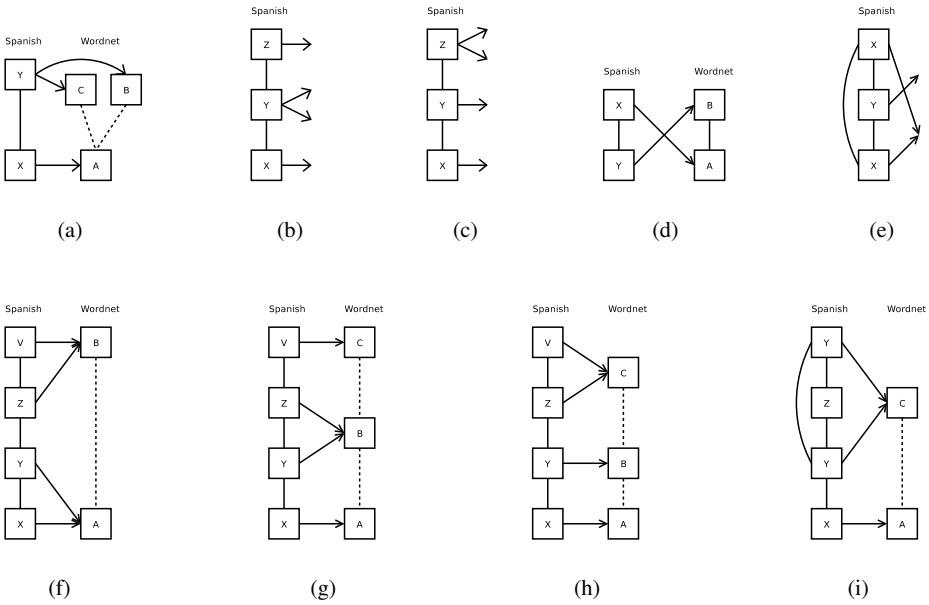
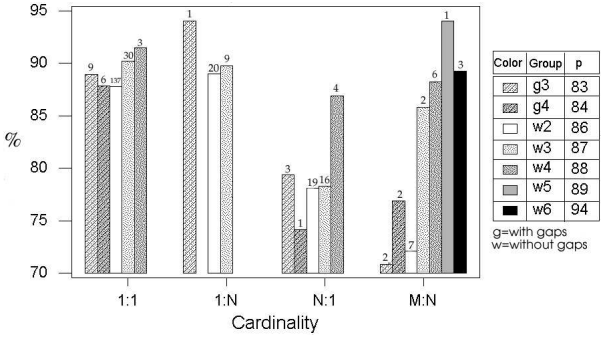


Table 2. Left: Cardinality sets. Right: Chains of cardinality sets.

Cardinality	Conn.	$\bar{p}$
1:1	184	0.88
1:N	30	0.89
N:1	49	0.79
M:N	23	0.81
crossing	11	



7 Results and Conclusions

A set of 1263 *StS* where induced from *WtS* following a simple heuristic, with a 96.7% estimated correctness percentage §4.

For the 1263 Spanish words, their branches were developed and their *StS* or *WtS* to Wordnet were identified. It was seen that developing chains with five ancestors is enough to get all the relevant information. So, given a *StS* c, the concept of *PRB*(c) was introduced in order to study the relationship between the Spanish taxonomy and Wordnet §5.

The internal structures of *PRB*s were studied. Depending on the level of the first ancestor with an association, different groups were obtained. *PRB*s with the first ancestor with an association at level 1 are faced in §6 where all possible patterns taking place in this family of *PRB*s are identified and displayed in figure 2.

Finally the research extracted three factors that affect relationship between the structure of the *PRB* and the logistic probability of the base *StS* obtained with the model previously developed in [Farr02]: the cardinality of the relation between the branches of the *PRB*, the number of Spanish senses with an association in the *PRB* and the existence of gaps in the Spanish branch, summarizing the results in table 2.

The main result of this paper is that, indeed, the probability of the *StS* c used to generate the *PRB* tends to increase mainly with the number of Spanish senses with an association in the *PRB*. That is, in fact, a quite surprising and interesting result since the logistic model was based on the solution sets of methods which don't use the Spanish taxonomy at all.

8 Future Work

After analyzing the simplest case, some parameters that affect the probabilities of the *StS* used to generate *PRB*s have been identified. Research is now centered on monosemic Spanish words with several association in order to evaluate how these parameters help to choose the correct associations, and what other factors appear that were not detected during the present study. Plans are in progress for analyzing the cases of polysemic words with only one link.

When all the factors would have arisen after the preliminary studies pointed above, the work will be centered on how to take profit of the taxonomic relation, and how to infer data from upper levels of *PRB*s.

References

Atse98. J. Atserias, S. Climent, J. Farreres, G. Rigau, and H. Rodríguez. *Recent Advances in Natural Language Processing*, chapter Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. Jon Benjamins Publishing Company, Amsterdam, The Netherlands, 1998.

Farr02. J. Farreres, K. Gibert, and H. Rodríguez. Semiautomatic creation of taxonomies. In G. Ngai *et al.*, editor, *Proceedings of SEMANET'02*, Taipei, 2002.

Farr98. J. Farreres, G. Rigau, and H. Rodríguez. Using wornet for building wordnets. In *Proceedings of COLING/ACL "Workshop on Usage of WordNet in Natural Language Processing Systems"*, Canada, 1998.

Rigau98. G. Rigau. *Automatic Acquisition of Lexical Knowledge from MRDs*. PhD thesis, Universitat Politècnica de Catalunya, 1998.