Semiautomatic creation of taxonomies

Abstract

In this paper we face the automatic construction of a lexical taxonomy for the Spanish language using as input a taxonomy of English (WordNet)[1] and a set of bilingual (English/Spanish) resources.

Although applied to Spanish/English our method claims to be general enough to be applied in the cases a skeletal taxonomy already exists and we dispose of methods mapping items to this taxonomy with known confidence scores.

1 Introduction

The automatic construction of accurate taxonomies from sets of incomplete, partially overlapping knowledge sources with different coverage/confidence characteristics has been object of interest for many researchers(Bisson et al., 2000; Faatz et al., 2001; Mihalcea and Moldovan, 2001).

The case of lexical taxonomies is specially challenging because of their huge size that advises against a manual construction.

Using bilingual dictionaries for mapping words or senses of a language to English counterparts is not new (Caraballo, 2001). In fact the work presented here can be considered an extension of (Author1, 1997; Author1, 1998) that was the base of the approach to building the Spanish WordNet within EuroWordNet[2] project.

From then within our group several efforts have been devoted to:

1. validating and debugging the content of the most important (higher levels in the hierarchy) and less confident (lesser scoring) synsets
2. extending wordnet in several tematic domains

Our approach starts from the acceptation of WordNet structure as a skeleton where words of the language to be studied can be placed[3]. It must be pointed out that this work is centered only in the nominal part of WN.

WordNet is a wide-coverage lexical-conceptual taxonomy of English. Its units (synsets) group together a set of words (variants) related with a loose form of synonymy. Synsets can be related by several semantic relations, being hypernymy/hyponymy the most important.

In the work we are presenting the statistical base is studied in depth and formalized.

2 Extracting candidate tuples

The core of the original method was the extraction of a huge amount of Spanish/English word pairs, Hbil, from both parts of a bilingual dictionary(Alvar, 1987).

Then 17 automatic methods were constructed that, from Hbil, generated 17 sets of connections between Spanish words and WN synsets. These methods followed different criteria of pairing in such a way that that resulting sets presented diverse degrees of overlapping and different degrees of quality measured in terms of precision and coverage. The 17 automatic methods can be grouped as shown below:

- methods 1, 2, 3, 4, 5, 6, 7, 8: connect Spanish words to WN synsets taking into account the equivalence between Spanish and English words (1:1, 1:N, M:1, M:N) and the polysemy of English words in WN (mono, poly).
- methods 9, 10, 11, 12: in case a Spanish word has more than one translation, connections are extracted if the translations were linked by a taxonomic relation in WN

[1] http://www.cogsci.princeton.edu/~wn/

[2] http://www.hum.uva.nl/~ewn/

[3] minor changes on semantic relationships are considered too.

Table 1: Method Evaluation

method	set volume	accuracy
1	3704	92%
2	935	89%
3	1888	89%
4	2690	85%
5	5123	80%
6	1450	75%
7	11687	58%
8	40299	61%
9	1256	79%
10	1432	51%
11	2202	57%
12	1846	60%
13	23829	56%
14	24740	61%
15	4567	75%
16	3164	85%
17	510	78%

- methods 13, 14, 15: they take profit of the semantic relations in WN using it as a semantic space for measuring conceptual distance between different elements (English translations of Spanish words occurring in definitions of a head-word entry, English translations of a head-word and its genus, and so).See(E. and G, 1995)
- method 16: it generates connections taking profit of the coincidence of English words from the same translation in the same synset.
- method 17: it generates connections taking profit of thematic marks in the bilingual dictionary.

Each of these methods generates a list of pairs of the kind synset-word. Each connection can be generated by multiple methods.

3 Selecting the most appropriate tuples: first approach

In order to evaluate the quality of the different methods, a random stratified sample of around 2,500 links covering all the sets was extracted and verified manually. The results of this evaluation are presented in table 1.

All of the methods that gave a percentage of correctness of 85% or better were selected to build what was called *Subset1*. This gave as a result an initial set of 10.982 connections between Spanish words and WN synsets from a potential volume of 66,609.

Based on the supposition that, if a connection is a joint solution of two methods, its probability to be correct would be higher, and then the joint evaluation of both methods will be higher that that of each method separately, and having checked the high degree of intersection between solutions of the different methods, we decided to add to the previous set of connections (*Subset1*) those connections occurring as simultaneous solution of two of the methods not considered in the previous phase.

We decided to preselect the promising intersections. This way, the synsets belonging to those intersections with accuracy above 85% would be included, leaving apart the intersections with higher error. We decided to study only the solutions from intersections with 85% or above.

The links of those methods not selected for Subset1 were crossed, and the volume of each intersection set was calculated. The percentage of correct solutions in each set can be understood as an estimation of the probability that one connection proposed by some pair of methods is correct. The solutions of certain pairs of methods presented accuracies equal or above 85%. But some didn't have a sample significant enough to ensure a reliable estimation of the probability of producing correct solutions.

In this second step of the sampling we proceeded to complete the manual evaluation of those groups that seem promising. The results are summarized in table 2. The table identified 14 intersections with an accuracy equal or above 85% (in bold in the table). All connections belonging to those cells were selected to form *Subset2*. *Subset2* was formed, then, by:

- All links proposed by methods m1-m4,m16 (10,982 connections) which had been accepted in *Subset1*.
- All links proposed by m11xm5, m11xm6, ... , m6xm8 (bold cells) adding up to 7,244 connections.

Table 3 shows the comparison of volumes and accuracies between the first sample stage (*Subset1*), the connections extracted in the second stage of sampling (*Intersections*) and the set resulting from the fusion of both (*Subset2*). The cardinality of the set is less than the addition of cardinalities because there

Table 2: Intersections

vol(%)	m13	m14	m15	m12	m10	m5	m6	m7	m8
m11	855(70)	828(71)	435(79)	449(58)	405(6)	**76(86)**	**107(89)**	0	1872(67)
m13		15736(79)	**1849(85)**	576(68)	419(71)	**2076(86)**	**556(86)**	3146(72)	15105(64)
m14			**2041(86)**	571(71)	428(72%	**2536(88)**	**592(86)**	3777(75)	13246(67)
m15				391(79)	325(80)	**205(95)**	**180(95)**	**215(100)**	3114(77)
m12					1432(67)	69(78)	68(7)	0	1463(65)
m10						69(77)	61(70)	0	1101(67)
m5							0	**77(100)**	**178(88)**
m6								28(77)	**78(96)**

Size and accuracy of each intersection is presented

Table 3: Intersection comparison

Set	#Links	#Synsets	#Words	%
Subset1	10982	7131	8396	87.4
Inters	7244	5852	3939	85.6
Subset2	15535	10786	9986	86.4

are connections belonging to both sets. This gives an idea that the degree of intersection is far greater than 2, what is worth studying. This is the origin of the present work.

4 Extending the coverage

Spanish WordNet has been further developed, adding new synsets and variants and validating manually the links in *Subset2*. As a result we now dispose of a wider and more accurate database that allows us to perform more robust estimations of confidence score factors for the different methods. We will call this manually verified database *Subset3*.

The result is that now, from the manually validated links, 20,013 correspond to connections extracted from automatic methods. The difference between these two figures (15,535 and 20,013) is due to the insertions performed during the manual evaluation process of *Subset3*. Those connections, as has already been pointed out, don't occur in *Subset2* but can belong to the set of results of some method not selected so far.

We decided to take as study population all of the validations obtained in any of the different phases (20,013) and to construct with those connections a third set of connections, all of them manually evaluated as correct(OK) or incorrect(KO), which will be used as test to evaluate again the whole population, and try to obtain, by means of a more detailed study of intersections, a *Subset4* which enhances the existing results in *Subset3*.

From the 20,013 evaluations there are 17,140 correct and 2,873 incorrect, giving an accuracy ratio over 85%. This result some way validates the previous work, as it was our intention to obtain a large set of connections with a value of above 85%.

In obtaining of *Subset2* it was evident the high degree of intersections between the different methods, a degree much larger than 2, but only the intersections of two methods were studied. It is our goal to study if the study of intersections could be exploited to extract from them an individual evaluation for each connection, and a formula that allows to calculate this value for new connections depending on the set of methods that propose them.

Concretely, the aim of the present work is to study the statistical behavior of the links regarding the set of methods which propose them as a solution, and not only intersections of two methods. To do this all the data has been condensed in a matrix of 66,609 vectors, one vector for each link, of the kind

link m_1 m_2 ... m_{16} m_{17} *eval*

where m_i are booleans indicating membership of the link to the set of solutions of method *i*, *eval* is the manual evaluation accepting one of two values (OK being correct,KO being incorrect), and *link* is the pair (WN1.5 synset,Spanish word). From this matrix the 20,013 rows with manual verification have been extracted to be studied separately, with the aim of obtaining some statistical measure that permits us to select the good connections of the set, in order to apply later the statistic to the whole population.

4.1 Descriptive analysis

Using the set of 20.013 validated links, the accuracy of each method can be reevaluated in a more precise way than in the first stage of the sampling (table 1). The results of the reevaluation are shown in table 4.

Comparing both tables the methods have different accuracies. While some methods (1, 15, 16) keep a similar accuracy, some suffer a light decrement (2, 3, 4), and most of them have a significant

Table 4: Accuracy reevaluation

method	correct	incorrect	mean
1	3066	294	91.25%
2	661	104	86.40%
3	1173	394	74.85%
4	1348	598	69.27%
5	3404	311	91.62%
6	759	131	85.28%
7	2099	170	92.50%
8	5686	998	85.06%
9	807	174	82.26%
10	413	130	76.05%
11	733	214	77.40%
12	506	156	76.43%
13	8063	1158	87.44%
14	10264	1402	87.98%
15	1917	585	76.61%
16	2290	364	86.28%
17	351	40	89.76%

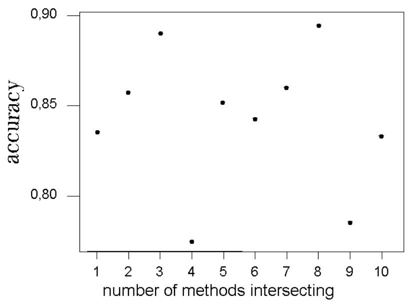
increment (5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17). But these variations must be studied with caution.

On one side, methods 1, 2, 3, 4 and 16 were accepted completely in the first stage of the sampling. Thus for these methods it is fair to study the changes observed. Concretely methods 1 and 16 maintain the same levels, while methods 2, 3 and 4 suffer a light decrease, more important in the cases of methods 3 and 4.

The rest of the methods were not selected completely, but based on intersections with other methods, and must be analyzed from another point of view. The general increment in the rest of methods comes from the manually added connections during revision process.

Another assumption taken in the generation of the set was that the accuracy increased with the number of methods which produced them. In figure 1 the percentage of correct links by the number of coincident methods is shown, and it proves the assumption is not be ensured. It doesn't seem to exist a correlation between the number of methods that produce a connection and the percentage of correct solutions. Figure 1 depicts just sets of methods intersecting, but they don't distinguish which of the methods are intersecting in each case. May be there are some methods of higher predictive quality than others, and this should be taken into account. So, it seems better to analyze the relationship between those methods according to the coincident links they provide.

Figure 1: dot plot



If a study is performed using the vectors previously shown, each connection is being evaluated against the set of methods that produce it. In this analysis, the fact that in the second stage of the sampling the less promising sets of solutions of certain pairs of methods were infra-represented doesn't suppose any risk, as now it is going to be considered the fact that each connection is correct or not in relation to the total set of methods that produce it.

4.2 Logistic regression model

We have applied without any success the Principal Components Analysis method, trying to find a spatial dispersion that separated correct and incorrect connections. We chose then to apply another statistical method more appropriate to the problem: logistic regression, which is used to obtain an adjustment for the accuracy of a connection from the set of methods that propose it.

4.2.1 Method formalization

The set of boolean variables $\mathrm{m}_1 ... \mathrm{m}_{17}$ defines 2^{17} possible combinations of methods that propose a certain connection. Associating to each connection a vector of this kind, the same description will be used for all connections proposed by the same group of methods and they will collapse in the same point. Thus a new matrix can be constructed that for the 2^{17} possibilities keeps the number of correct evaluations and the total of evaluations, being the number of incorrect ones the difference between both values:

m_1 m_2 ... m_{16} m_{17} *nok ntot*

where m_i are boolean values indicating the membership of a given link to the set of solutions of method *i*, *nok* is the number of correct evaluations for the set of solutions of every group of methods, and *ntot* accumulates the total number of evaluations. It is clear that the probability that a link would be correct can be estimated by the following expression $P(OK) = NOK/NTOT$ (considering that the probability is the limit of the relative frequency). The logistic regression is a technique which allows finding a model (in the mathematical sense) for approximating $P(OK)$ on the basis of a set of explicative variables (in our case m_1, m_2, ... m_{17}).

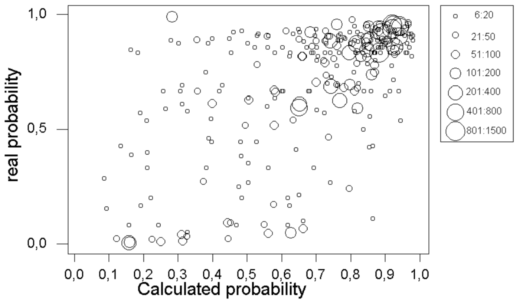
In order to fit the formula $\log(P(nok/ntot)) = \beta_0 + \beta_1 m_1 + ... + \beta_{17} m_{17}$, the least square criteria is used for every value of β_i. It was observed that the analysis is disturbed by a series of combinations of methods with very low frequency in the sample (with very low *ntot*), and it was decided to restrict (in this first phase) the study to those combinations of methods more represented in the sample. Then the analysis was redone for those rows with $ntot > 5$. This elimination supposes the loss of 1,210 evaluations, which mean a 10% of the total number, leaving a set of 20,934.

4.2.2 Backward method

The results can be enhanced. The P value of each of the factors shows which methods are significant for explaining the probability of a link of being OK. For a P value lower than 0.05, the method is significant in the model; this means that to be or not a result of this method is influencing the probability of OK for a given link. The objective should be to find the minimal set of significant methods. But this is an exponential problem, as all combinations should be tested. The backward method is useful to find a local optimum iteratively deleting the less informative variable between the non significant ones. In the case of the present study, methods m_3 and m_4 have a P value very close to 1, meaning they are not significant. Performing the backward method both of them would be eliminated. All P-values for the remaining methods are close to 0 and the Pearson goodness-of-fit test gives a P-value of 0.000, meaning the model and all its methods are significant.

To evaluate the diagonal tendency of the model, a plot has been done clustering the points in seven ranked groups, and showing each point with a circle of a diameter equivalent to the size of the group. It is shown in figure 2. In the graph the diagonal tendency can be clearly seen showing there is a correlation between both measures, the calculated and the real ones. This is a good indicator, as it means that the method approximates quite correctly the problem. But, what is more important, is that the groups with the larger sizes are positioned more to the diagonal center than the ones poorly evaluated. The points have been represented so that the more populated ones are drawn larger than the smaller ones.

Figure 2: Final dispersion of points clustered by size The graphic shows on the vertical axis the real percentage of OK values, and on the horizontal axis the fitted prob(OK).



5 Selecting the most appropriate tuples: final approach

After selecting only the methods that are significant to the model, there is a general formula to calculate the % from the results of the regression model: p(OK)=$\frac{e^{\sum m_i k_i}}{1+e^{\sum m_i k_i}}$ where the values m_i correspond to table 5.

The formula obtained was applied to the total set of 66,604 connections proposed by the automatic methods, in order to obtain an individual evaluation for each of the points. By means of this estimation of the probability of correctness of each point, a global percentage can be estimated for the whole population. In this case it has given 84.86%. We dispose now of an estimation of goodness for any of the proposed connections. We can, thus, study some thresholding mechanism for deciding whether to accept or not a given connection. The results are presented in table 6. For each threshold the number of total connections selected has been calculated, the number of synsets and Spanish words related, the number of OK and KO evaluations, the estimation of the real global percentage($OK/(OK + KO)$), the global estimation from the formula, and the degree of polysemy

Table 6: Subset4 thresholds

threshold	#links	#synsets	#spwords	NOK	NKO	Real % Est	% Est	Poly
none	66609	19378	13336	17140	2873	85.64	84.86	4.99
0.1	66597	19378	13336	17133	2873	85.63	84.87	4.99
0.2	66538	19378	13333	17090	2867	85.63	84.93	4.99
0.3	66438	19377	13329	17025	2847	85.67	85.02	4.98
0.4	65995	19365	13317	16784	2765	85.84	85.34	4.95
0.5	65710	19346	13302	16660	2726	85.93	85.51	4.93
0.6	64876	19254	13205	16277	2566	86.38	85.90	4.91
0.7	61435	18802	12399	15106	2056	88.02	87.01	4.95
0.8	56865	17424	11375	13031	1494	89.71	87.91	4.99
0.82	50695	17254	11338	12206	1289	90.48	88.66	4.47
0.84	50177	17037	11288	11944	1242	90.58	88.72	4.44
0.85	49279	16535	11175	11320	1136	90.87	88.80	4.40
0.855	47831	16477	11141	11078	1105	90.97	88.90	4.29
0.858888	29664	15319	10687	9887	940	91.31	90.76	2.77
0.86	29644	15317	10683	9876	934	91.35	90.76	2.77
0.87	29128	15065	10587	9616	858	91.80	90.83	2.75
0.88	29043	15001	10559	9555	849	91.83	90.85	2.75
0.89	19202	10975	8491	7161	567	92.68	91.99	2.26
0.90	19041	10936	8482	7102	554	92.75	92.02	2.24
0.91	11840	8933	7528	5656	413	93.19	93.03	1.57
0.92	11770	8891	7506	5606	405	93.26	93.04	1.56
0.925	7280	5609	5109	2790	248	91.83	93.55	1.42
0.93	3801	3491	3156	2261	163	93.27	94.32	1.20

Table 5: Factors

m_i	k_i
m_0	0.62426
m_1	1.39341
m_2	1.2
m_5	1.39876
m_6	1.39876
m_7	1.44022
m_8	1.18182
m_9	0.8399
m_{10}	0.7417
m_{11}	-0.65116
m_{12}	-0.7874
m_{13}	-0.30914
m_{14}	0.73912
m_{15}	-1.03721
m_{16}	-0.53728
m_{17}	-2.29611

of the Spanish words ($\#spwords/\#synsets$).

From the results of table 6 several issues can be observed. On the one side, in the initial set of connections there is an index of polysemy of 4.99, when the global polysemy of Spanish nouns is 1.8, see(Rigau, 1998). This value would correspond to a threshold between 0.90 and 0.91. If we accept senses from Spanish to be translated to more than one sense in WN, a higher polysemy degree could be accepted, but never reaching 4.99.

It is interesting to note the decreasing evolution of the illustrative variable *polysemy* as the threshold increases, reaching at the limit a value of 1.2, very close to monosemic values. In order to choose a threshold dividing the population between good and bad results, the polysemy degree of the total population is a good indicator, and surprisingly well correlated with the formula.

Observing the row at 0.858888% a drastic decrement of the number of connections occurs, but not followed by the number of words or synsets. it means, only the polysemy degree is being decremented to 2.77 without a great loss of representativity of the whole set, which maintains a global estimation of 90.76%. This decrement is caused by the disappearance of nearly 16.000 connections uniquely generated by method 8. The second jump

Table 7: Subset4 comparison

Set	#Links	#Synsets	#Words	%
Subset3	54753	38172	41129	98.99
Regress	29664	15319	10687	91.3
Subset4	74516	40334	42647	94.80
Increm	19763	2162	1518	unkown

takes place in the row of 91%, where losing a significant number of words and synsets, the polysemy lowers to 1.57 with a global estimation of 93.03%. Beyond this point the loss in coverage is much greater than the gain in precision.

What is really interesting is that, independently of the global estimation of the set, each connection has its own estimation. Then, if a threshold of 86% is chosen, all of the 29,644 connections chosen have each its own estimation, and this is an important information; a verification can be designed depending on the estimation of the points, for example, starting with the lower ones, to eliminate a larger number of errors, or starting with the higher ones, to ensure a higher degree of correct results.

In a similar way, if the threshold of 86% was selected, from the 2.653 Spanish words lost, the connection with the highest estimation could be included, which will always be lower than 86%, trying to rescue the largest number of words possible, placing into the set at least one sense for those words. Or senses for those words could be discarded from upper to lower until one correct sense is found.

This means that, once this result has been obtained, there is a wide number of possible ways and solutions to optimize the effort made, and to maximize the number of correct solutions per manual evaluation.

Anyway a last comment must be said to indicate that this process doesn't help to avoid the need for a final work of hand evaluating the different connections obtained in order to delete the wrong results that have been included in the final set.

In order to give a final result, and to compare it with the results previously obtained, a threshold of 0.858888% was selected as cut point, and all the results above this limit were accepted. Table 7 shows the results. The correctness of the increment is unknown but falls into the limits of *Subset4*, and the correctness of both the regression and Subset4 are only estimated.

6 Conclusions and future work

A method for automatically linking Spanish words to a skeletal lexical-conceptual taxonomy has been presented. The system uses WordNet as base for linking to it Spanish words obtained by a set of different methods from a set of bilingual sources.

A sound statistical base has been tested for scoring the different candidates and a threshold mechanism has been proposed.

The system clearly outperforms our previous approach that was on the base of building the Spanish WordNet.

The system is claimed to be general enough to be applied to other languages.

References

M. Alvar, editor. 1987. *Diccionario General Ilustrado de la Lengua Española VOX*. Biblograf S.A, Barcelona, Spain.

Author1. 1997.

Author1. 1998.

G. Bisson, C. Nedellec, and D. Cañamero. 2000. Designing clustering methods for ontology building - the mo'k workbench. *ECAI*.

Sharon A. Caraballo. 2001. *Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text*. Ph.D. thesis, Brown University.

Agirre E. and Rigau G. 1995. A proposal for word sense disambiguation using conceptual distance. *Recent Advances in Natural Language Processing*.

A. Faatz, S. Hormann, C. Seeberg, and R. Steinmetz. 2001. Conceptual enrichment of ontologies by means of a generic and configurable approach. *ESSLII*.

Rada Mihalcea and Dan I. Moldovan. 2001. Automatic generation of a coarse grained wordnet. *NAACL 2001 Workshop on WordNet and Other Lexical Resources*.

G. Rigau. 1998. *Automatic Acquisition of Lexical Knowledge from MRDs*. Ph.D. thesis, Polytechnic University of Catalonia.