Towards a Semantic Classification of Spanish Verbs Based on Subcategorisation Information

Eva Esteve Ferrer
Department of Informatics
University of Sussex
Brighton, BN1 9QH, UK
E.Esteve-Ferrer@sussex.ac.uk

Abstract

We present experiments aiming at an automatic classification of Spanish verbs into lexical semantic classes. We apply well-known techniques that have been developed for the English language to Spanish, proving that empirical methods can be re-used through languages without substantial changes in the methodology. Our results on subcategorisation acquisition compare favourably to the state of the art for English. For the verb classification task, we use a hierarchical clustering algorithm, and we compare the output clusters to a manually constructed classification.

1 Introduction

Lexical semantic classes group together words that have a similar meaning. Knowledge about verbs is especially important, since verbs are the primary means of structuring and conveying meaning in sentences. Manually built semantic classifications of English verbs have been used for different applications such as machine translation (Dorr, 1997), verb subcategorisation acquisition (Korhonen, 2002a) or parsing (Schneider, 2003). (Levin, 1993) has established a large-scale classification of English verbs based on the hypothesis that *the meaning of a verb and its syntactic behaviour are related, and therefore semantic information can be induced from the syntactic behaviour of the verb*. A classification of Spanish verbs based on the same hypothesis has been developed by (Vázquez et al., 2000). But manually constructing large-scale verb classifications is a labour-intensive task. For this reason, various methods for automatically classifying verbs using machine learning techniques have been attempted ((Merlo and Stevenson, 2001), (Stevenson and Joanis, 2003), (Schulte im Walde, 2003)).

In this article we present experiments aiming at automatically classifying Spanish verbs into lexical semantic classes based on their subcategorisation frames. We adopt the idea that a description of verbs in terms of their syntactic behaviour is useful for acquiring their semantic properties. The classification task at hand is achieved through a process that requires different steps: we first extract from a partially parsed corpus the probabilities of the subcategorisation frames for each verb. Then, the acquired probabilities are used as features describing the verbs and given as input to an unsupervised classification algorithm that clusters together the verbs according to the similarity of their descriptions. For the task of acquiring verb subcategorisation frames, we adapt to the specificities of the Spanish language well-known techniques that have been developed for English, and our results compare favourably to the sate of the art results obtained for English (Korhonen, 2002b). For the verb classification task, we use a hierarchical clustering algorithm, and we compare the output clusters to a manually constructed classification developed by (Vázquez et al., 2000).

2 Acquisition of Spanish Subcategorisation Frames

Subcategorisation frames encode the information of how many arguments are required by the verb, and of what syntactic type. Acquiring the subcategorization frames for a verb involves, in the first place, distinguishing which constituents are its arguments and which are adjuncts, elements that give an additional piece of information to the sentence. Moreover, sentences contain other constituents that are not included in the subcategorisation frames of verbs: these are sub-constituents that are not structurally attached to the verb, but to other constituents.

2.1 Methodology and Materials

We experiment our methodology on two corpora of different sizes, both consisting of Spanish newswire text: a 3 million word corpus, hereafter called small corpus, and a 50 million word corpus, hereafter called large corpus. They are both POS tagged and partially parsed using the MS-analyzer, a partial parser for Spanish that includes named entities recognition (Atserias et al., 1998).

In order to collect the frequency distributions of Spanish subcategorisation frames, we adapt a methodology that has been developed for English to the specificities of the Spanish language ((Brent, 1993), (Manning, 1993), (Korhonen, 2002b)). It consists in extracting from the corpus pairs made of a verb and its co-occurring constituents that are a possible pattern of a frame, and then filtering out the patterns that do not have a probability of co-occurrence with the verb high enough to be considered its arguments.

We establish a set of 11 possible Spanish subcategorisation frames. These are the plausible combinations of a maximum of 2 of the following constituents: nominal phrases, prepositional phrases, temporal sentential clauses, gerundive sentential clauses, infinitival sentential clauses, and infinitival sentential clauses introduced by a preposition. The individual prepositions are also taken into account as part of the subcategorisation frame types.

Adapting a methodology that has been thought for English presents a few problems, because English is a language with a strong word order constraint, while in Spanish the order of constituents is freer. Although the unmarked order of constituents is Subject Verb Object with the direct object preceding the indirect object, in naturally occurring language the constituents can be moved to non-canonical positions. Since we extract the patterns from a partially parsed corpus, which has no information on the attachment or grammatical function of the constituents, we have to take into account that the extraction is an approximation. There are various phenomena that can lead us to an erroneous extraction of the constituents. As an illustrative example, in Spanish it is possible to have an inversion in the order of the objects, as can be observed in sentence (1), where the indirect object *a Straw* ("to Straw") precedes the direct object *los alegatos* ("the pleas").

> (1) El gobierno chileno presentará hoy a Straw los alegatos (. . .).
>
> "The Chilean government will present today to Straw the pleas (. . .)".

Dealing with this kind of phenomenon introduces some noise in the data. Matching a pattern for a subcategorisation frame from sentence (1), for example, we would misleadingly induce the pattern [_ PP(a)] for the verb *presentar*, "present", when in fact the correct pattern for this sentence is [_ NP PP(a)].

The solution we adopt for dealing with the variations in the order of constituents is to take into account the functional information provided by clitics. Clitics are unstressed pronouns that refer to an antecedent in the discourse. In Spanish, clitic pronouns can only refer to the subject, the direct object, or the indirect object of the verb, and they can in most cases be disambiguated taking into account their agreement (in person, number and gender) with the verb. When we find a clitic pronoun in a sentence, we know that an argument position is already filled by it, and the rest of the constituents that are candidates for the position are either discarded or moved to another position. Sentence (2) shows an example of how the presence of clitic pronouns allows us to transform the patterns extracted. The sentence would normally match with the frame pattern [_ PP(por)], but the presence of the clitic (which has the form *le*) allows us to deduce that the sentence contains an indirect object, realised in the subcategorisation pattern with a prepositional phrase headed by *a* in second position. Therefore, we look for the following nominal phrase, *la aparición del cadáver*, to fill the slot of the direct object, that otherwise would have not been included in the pattern.

> (2) Por la tarde, agentes del cuerpo nacional de policía **le** comunicaron por teléfono la aparición del cadáver.
>
> "In the afternoon, agents of the national police **clitic_IO** reported by phone the apparition of the corpse.".

The collection of pairs *verb + pattern* obtained with the method described in the last section needs to be filtered out, because we may have extracted constituents that are in fact adjuncts, or elements that are not attached to the verb, or errors in the extraction process. We filter out the spurious patterns with a Maximum Likelihood Estimate (MLE), a method proposed by (Korhonen, 2002b) for this task. MLE is calculated as the ratio of the frequency of $pattern_i$+$verb_j$ over the frequency of $verb_j$. Pairs of *verb+pattern* that do not have a probability of co-occurring together higher than a certain threshold are filtered out. The threshold is determined empirically using held-out data (20% of the total of the corpus), by choosing from a range of values between 0.02 and 0.1 the value that yields better results against a held-out gold standard of 10 verbs. In our experiments, this method yields a threshold value of 0.05.

2.2 Experimental Evaluation

We evaluate the obtained subcategorisation frames in terms of precision and recall compared to a gold

	No Prep. Groups			Preposition Groups		
Corpus	Prec	Rec	F	Prec	Rec	F
Small	65	62	63	63	61	62
Baseline	25	78	38	31	82	45
Large	70	60	65	71	61	**66**
Baseline	8	96	14	8	96	14

Table 1: Results for the acquisition of subcategorisation frames.

standard. The gold standard is manually constructed for a sample of 41 verbs. The verb sample is chosen randomly from our data with the condition that both frequent and infrequent verbs are represented, and that we have examples of all our subcategorisation frame types. We perform experiments on two corpora of different sizes, expecting that the differences in the results will show that a large amount of data does significantly improve the performance of any given system without any changes in the methodology. After the extraction process, the small corpus consists of 58493 pairs of *verb+pattern*, while the large corpus contains 1253188 pairs.[1] Since we include in our patterns the heads of the prepositional phrases, the corpora contain a large number of pattern types (838 in the small corpora, and 2099 in the large corpora). We investigate grouping semantically equivalent prepositions together, in order to reduce the number of pattern types, and therefore increment the probabilities on the patterns. The preposition groups are established manually.

Table 1 shows the average results obtained on the two different corpora for the 41 test verbs. The baselines are established by considering all the frame patterns obtained in the extraction process as correct frames. The experiments on the large corpus give better results than the ones on the small one, and grouping similar prepositions together is useful only on the large corpus. This is probably due to the fact that the small corpus does not suffer from a too large number of frame types, and the effect of the groupings cannot be noticed. The F measure value of 66% reported on the third line of table 1, obtained on the large corpus with preposition groups, compares favourably to the results reported on (Korhonen, 2002b) for a similar experiment on English subcategorization frames, in which an F measure of 65.2 is achieved.

3 Clustering Verbs into Classes

We use a bottom-up hierarchical clustering algorithm to group together 514 verbs into K classes. The algorithm starts by finding the similarities between all the possible pairs of objects in the data according to a similarity measure S. After having established the distance between all the pairs, it links together the closest pairs of objects by a linkage method L, forming a binary cluster. The linking process is repeated iteratively over the newly created clusters until all the objects are grouped into one cluster. K, S and L are parameters that can be set for the clustering. For the similarity measure S, we choose the Euclidean distance. For the linkage method L, we choose the *Ward* linkage method (Ward, 1963). Our choice of the parameter settings is motivated by the work of (Stevenson and Joanis, 2003). Applying a clustering method to the verbs in our data, we expect to find a natural division of the data that will be in accordance with the classification of verbs that we have set as our target classification. We perform different experiments with different values for K in order to test which of the different granularities yields better results.

3.1 The Target Classification

In order to be able to evaluate the clusters output by the algorithm, we need to establish a manual classification of sample verbs. We assume the manual classification of Spanish verbs developed by (Vázquez et al., 2000). In their classification, verbs are organised on the basis of meaning components, diathesis alternations and event structure. They classify a large number of verbs into three main classes (Trajectory, Change and Attitude) that are further subdivided into a total of 31 subclasses. Their classification follows the same basic hypotheses as Levin's, but the resulting classes differ in some important aspects. For example, the Trajectory class groups together Levin's Verbs of Motion (move), Verbs of Communication (tell) and verbs of Change of Possession (give), among others. Their justification for this grouping is that all the verbs in this class have a Trajectory meaning component, and that they all undergo the Underspecification alternation (in Levin's terminology, the Locative Preposition Drop and the Unspecified Object alternations). The size of the classes at the lower level of the classification hierarchy varies from 2 to 176.

3.2 Materials

The input to the algorithm is a description of each of the verbs in the form of a vector containing the

[1] In all experiments, we post-process the data by eliminating prepositional constituents in the second position of the pattern that are introduced with the preposition *de*, "of". This is motivated by the observation that in 96.8% of the cases this preposition is attached to the preceding constituent, and not to the verb.

probabilities of their subcategorisation frames. We obtain the subcategorisation frames with the method described in the previous section that gave better results: using the large corpus, and reducing the number of frame types by merging individual prepositions into groups. In order to reduce the number of frame types still further, we only take into account the ones that occur more than 10 times in the corpus. In this way, we have a set of 66 frame types. Moreover, for the purpose of the classification task, the subcategorisation frames are enhanced with extra information that is intended to reflect properties of the verbs that are relevant for the target classification. The target classification is based on three aspects of the verb properties: meaning components, diathesis alternations, and event structure, but the information provided by subcategorisation frames only reflects on the second of them. We expect to provide some information on the meaning components participating in the action by taking into account whether subjects and direct objects are recognised by the partial parser as named entities. Then, the possible labels for these constituents are "no_NE", "persons", "locations", and "institutions". We introduce this new feature by splitting the probability mass of each frame among the possible labels, according to their frequencies. Now, we have a total of 97 features for each verb of our sample.

3.3 Clustering Evaluation

Evaluating the results of a clustering experiment is a complex task because ideally we would like the output to fulfil different goals. One the one hand, the clusters obtained should reflect a good partition of the data, yielding consistent clusters. On the other hand, the partition of the data obtained should be as similar as possible to the manually constructed classification, the gold standard. We use the Silhouette measure (Kaufman and Rousseeuw, 1990) as an indication of the consistency of the obtained clusters, regardless of the division of the data in the gold standard. For each clustering experiment, we calculate the mean of the silhouette value of all the data points, in order to get an indication of the overall quality of the clusters created. The main difficulty in evaluating unsupervised classification tasks against a gold standard lies in the fact that the class labels of the obtained clusters are unknown. Therefore, the evaluation is done according to the pairs of objects that the two groups have in common. (Schulte im Walde, 2003) reports that the evaluation method that is most appropriate to the task of unsupervised verb classification is the Adjusted Rand measure. It gives a value of 1 if the two classifications agree completely in which pairs of objects are clustered together and which are not, while complete disagreement between two classifications yields a value of -1.

No Named Entities			
Task	Mean Sil	Baseline	Radj
3-way	0.37	0	0.001
15-way	0.37	0	0.040
31-way	0.27	0	0.070

Table 2: Clustering evaluation for the experiment without Named Entities

Named Entities			
Task	Mean Sil	Baseline	Radj
3-way	0.37	0	0.01
15-way	0.31	0	0.07
31-way	0.22	0	0.03

Table 3: Clustering evaluation for the experiment with Named Entities

3.4 Experimental Results

We perform various clustering experiments in order to test, on the one hand, the usefulness of our enhanced subcategorisation frames. On the other hand, we intend to discover which is the natural partition of the data that best accommodates our target classification. The target classification is a hierarchy of three levels, each of them dividing the data into 3, 15, or 31 levels. For this reason, we experiment on 3, 15, and 31 desired output clusters, and evaluate them on each of the target classification levels, respectively.

Table 2 shows the evaluation results of the clustering experiment that takes as input bare subcategorisation frames. Table 3 shows the evaluation results of the experiment that includes named entity recognition in the features describing the verbs. In both tables, each line reports the results of a classification task. The average Silhouette measure is shown in the second column. We can observe that the best classification tasks in terms of the Silhouette measure are the 3-way and 15-way classifications. The baseline is calculated, for each task, as the average value of the Adjusted Rand measure for 100 random cluster assignations. Although all the tasks perform better than the baseline, the increase is so small that it is clear that some improvements have to be done on the experiments. According to the Adjusted Rand measure, the clustering algorithm seems to perform better in the tasks with a larger number of classes. On the other hand, the enhanced features are useful on the 15-way and 3-way classifications, but they are harmful in the 31-way classification. In spite of these results, a qualitative observation of the output clusters reveals that they are intuitively plausible, and that the evaluation is penalised by the fact that the target classes are of very different sizes. On the other hand, our data takes into account syntactic information, while the target classification is not only based on syntax, but also on other aspects of the properties of the verbs. These results compare poorly to the performance achieved by (Schulte im Walde, 2003), who obtains an Adjusted Rand measure of 0.15 in a similar task, in which she classifies 168 German verbs into 43 semantic verb classes. Nevertheless, our results are comparable to a subset of experiments reported in (Stevenson and Joanis, 2003), where they perform similar clustering experiments on English verbs based on a general description of verbs, obtaining average Adjusted Rand measures of 0.04 and 0.07.

4 Conclusions and Future Work

We have presented a series of experiments that use an unsupervised learning method to classify Spanish verbs into semantic classes based on subcategorisation information. We apply well-known techniques that have been developed for the English language to Spanish, confirming that empirical methods can be re-used through languages without substantial changes in the methodology. In the task of acquiring subcategorisation frames, we achieve state of the art results. On the contrary, the task of inducing semantic classes from syntactic information using a clustering algorithm leaves room for improvement. The future work for this task goes on two directions.

On the one hand, the theoretical basis of the manual verb classification suggests that, although the syntactic behaviour of verbs is an important criteria for a semantic classification, other properties of the verbs should be taken into account. Therefore, the description of verbs could be further enhanced with features that reflect on meaning components and event structure. The incorporation of name entity recognition in the experiments reported here is a first step in this direction, but it is probably a too sparse feature in the data to make any significant contributions. The event structure of predicates could be statistically approximated from text by grasping the aspect of the verb. The aspect of the verbs could, in turn, be approximated by developing features that would consider the usage of certain tenses, or the presence of certain types of adverbs that imply a restriction on the aspect of the verb. Adverbs such as ”suddenly”, ”continuously”, ”often”, or even adverbial sentences such as ”every day” give information on the event structure of predicates. As they are a closed class of words, a typology of adverbs could be established to approximate the event structure of the verb (Esteve Ferrer and Merlo, 2003).

On the other hand, an observation of the verb clusters output by the algorithm suggests that they are intuitively more plausible than what the evaluation measures indicate. For the purposes of possible applications, a hard clustering of verbs does not seem to be necessary, especially when even manually constructed classifications adopt arbitrary decisions and do not agree with each other: knowing which verbs are semantically similar to each other in a more “fuzzy” way might be even more useful. For this reason, a new approach could be envisaged for this task, in the direction of the work by (Weeds and Weir, 2003), by building rankings of similarity for each verb. For the purpose of evaluation, the gold standard classification could also be organised in the form of similarity rankings, based on the distance between the verbs in the hierarchy. Then, the rankings for each verb could be evaluated. The two directions appointed here, enriching the verb descriptions with new features that grasp other properties of the verbs, and envisaging a similarity ranking of verbs instead of a hard clustering, are the next steps to be taken for this work.

Acknowledgements

The realisation of this work was possible thanks to the funding of the Swiss FNRS project number 11-65328.01.

References

Jordi Atserias, Josep Carmona, Irene Castellón, Sergi Cervell, Montserrat Civit, Lluís Màrquez, M. Antonia Martí, Lluís Padró, Roser Placer, Horacio Rodríguez, Mariona Taulé, and Jordi Turmo. 1998. Morphosyntactic analysis and parsing of unrestricted spanish text. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, pages 1267–1272, Granada/Spain.

Michael Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Bonnie Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):1–55.

Eva Esteve Ferrer and Paola Merlo. 2003. Automatic classification of english verbs. Technical report, Université de Genève.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding Groups in Data - An Introduction to Cluster Analysis*. Probability and Mathematical Statistics. Jonh Wiley and Sons, Inc., New York.

Anna Korhonen. 2002a. Semantically motivated subcategorization acquisition. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon on Unsupervised Lexical Acquisition*, pages 51–58, Philadelphia,PA, July.

Anna Korhonen. 2002b. *Subcategorisation Acquisition*. Ph.D. thesis, University of Cambridge. distributed as UCAM-CL-TR-530.

Beth Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.

Christopher Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 235–242, Columbus/Ohio.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Gerold Schneider. 2003. A low-complexity, broad coverage probabilistic dependency parser for english. In *Proceedings of NAACL/HLT 2003 Student Session*, pages 31–36, Edmonton/Canada.

Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart. Published as AIMS Report 9(2).

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, page , Edmonton/Canada.

Gloria Vázquez, Ana Fernández, Irene Castellón, and M. Antonia Martí. 2000. Clasificación verbal: Alternancias de diátesis. *Quaderns de Sintagma. Universitat de Lleida*, 3.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, Sapporo/Japan.